NBER WORKING PAPER SERIES

U.S. SCHOOL FINANCE: RESOURCES AND OUTCOMES

Danielle V. Handel
Eric A. Hanushek

Working Paper 30769
http://www.nber.org/papers/w30769

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2022, Revised February 2023

We benefitted from very helpful comments and suggestions by Jay Greene, Larry Hedges, and Ludger Woessmann. This article will appear in Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, eds. forthcoming, 2023. Handbook of the Economics of Education. Vol. 7. Amsterdam: North Holland. Hanushek consults and testifies about school finance policy for both plaintiffs and defendants in education litigation in various states. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

U.S. School Finance: Resources and Outcomes
Danielle V. Handel and Eric A. Hanushek
NBER Working Paper No. 30769
December 2022, Revised February 2023
JEL No. H41,H72,I20,J08

ABSTRACT

The impact of school resources on student outcomes was first raised in the 1960s and has been controversial since then. This issue enters into the decision making on school finance in both legislatures and the courts. The historical research found little consistent or systematic relationship of spending and achievement, but this research frequently suffers from significant concerns about the underlying estimation strategies. More recent work has re-opened the fundamental resource-achievement relationship with more compelling analyses that offer stronger identification of resource impacts. A thorough review of existing studies, however, leads to conclusions similar to those in the historical work: how resources are used is key to the outcomes. At the same time, the research has not been successful at identifying mechanisms underlying successful use of resources or for ascertaining when added school investments are likely to be well-used. Direct investigations of alternative input policies (capital spending, reducing class size, and salary incentives for teachers) do not provide clear support for such specific policy initiatives in the United States.

Danielle V. Handel
SIEPR
Stanford University
Stanford, CA 94305
dvhandel@stanford.edu

Eric A. Hanushek
Hoover Institution
Stanford University
Stanford, CA 94305-6010
and NBER
hanushek@stanford.edu

A data appendix is available at http://www.nber.org/data-appendix/w30769

1 Introduction

School finance in the United States is complicated. The separate states are primarily responsible for funding and for policy, but the states delegate substantial funding and operational decisions to local governments. The federal government does enter into both finance and policy decisions, albeit in limited ways and in specific functional areas. But a major complicating factor coloring the overall financing picture is that the courts have entered directly into the decision making process and in some states have even assumed a dominant role in overall school financing decisions.

The research in this chapter directly relates to the outcomes of school finance policy developed by legislatures and courts. The research differs, however, from most other work in the economics of education because of the swiftness with which it enters into the decision making on school finance both in and out of the courts.

The path of both the court actions and the related research can be traced back to three books that appeared nearly simultaneously. Two books set out the initial legal case for more equitable funding of schools (Coons, Clune, and Sugarman (1970), Wise (1968)). These books focused on revenues for school districts and identified the wide disparities in school funding that followed significant reliance on local property taxes. They played a role in the development of early court cases related to school funding. But, before either of these books was published, the issue of the relationship between funding and student outcomes was raised in the "Coleman Report" (Coleman et al. (1966)). This early government report pioneered the use of statistical analysis to investigate how school resources related to student achievement and suggested that expenditure differences among schools were not very important in determining school outcomes. The contrast between the court focus on revenues and the parallel questions about the relevance of revenue variations for student outcomes has remained over the past half century.

Shaped by a lawsuit in California in 1968, the court challenges to state finance systems revolved around disparities in funding that arose from using the local property tax as a mainstay of funding, a system that produced inequitable school opportunities as measured by funding. Progressively, court challenges moved across the states, and they evolved from pure equity cases to expanded questions about the adequacy of funding to meet goals of high quality schools.

The focus of this chapter is the evidence on funding and achievement that enters into the school finance court cases and not the legal cases and decisions per se. The courts have influenced the level and distribution of school funding, but we are not interested in analyzing their separate role. The parallel research on how funding relates to outcomes sometimes involves these court cases in their role in generating variation in funding but is more general.

The most basic school finance question that continues to be discussed is whether just changing the budget constraint for schools leads to better student outcomes. Even asking such a question is strange, because the simplest microeconomic theory would dismiss it out of hand. The underlying issues in the school finance discussions, however, are not standard textbook problems because school budgets are produced with a range of institutional and regulatory constraints: the uses that any additional money can be put to are restricted by specific spending requirements, by state and federal laws, by local teacher contracts, and by a myriad of other limitations that might make best use of any funds in the local school district exceedingly difficult. Thus, this fundamental consideration of the

relationship of funding and achievement becomes an empirical question, one which has recently again become a heavily researched topic.

A large literature developed from the Coleman Report to address the question of how different school inputs including total resources affected student outcomes. This earliest production function study suggested that school resources had little to do with student performance, and it led to a wide range of studies that delved into the determinants of student outcomes. Because decision making on school funding – both in state legislatures and in the courts – naturally related to consideration of the role of funding in ensuring quality schooling, this literature had direct policy linkages from the outset even though it was not motivated directly by overall issues of school finance. But, the pace of this line of research into education production functions slowed noticeably by the 1980s and 1990s as new insights waned.

The more recent path of empirical research toward more convincing identification of program impacts revived research into the relevant line of school finance studies. The search for more credible empirical evidence on the impacts of various school and other inputs has led to a new literature. The prior production function evidence included a number of studies that would not meet current quality standards for empirical analysis. The recent evidence provides a new look at the longstanding issues of resources and outcomes while paying much greater attention to the identification of key policy parameters.

Interestingly, the overall empirical results of the recent analyses tend to mirror those from the older production function work. When the recent estimates of the key impact parameters are standardized, the wide variation in estimated effects becomes very apparent. While the specific focus of the older and newer lines of research has been somewhat different, both lines of research point to significant heterogeneity in the impact of resources on outcomes. The individual impact studies provide a range of specific impact parameters. Once these are put on a common scale, the heterogeneity of results underscores a necessity of focusing on how resources are used.

The variation in findings across studies also raises questions about what generalizations are appropriate from either line of research. Both the historic and more contemporary studies include a meaningful proportion of estimated resource parameters that are not statistically significant. Moreover, there is currently little explanation about the mechanisms underlying these widely different point estimates for the impact of resources.

This discussion begins with an overview of the structure of funding for U.S. education followed by a description of the pattern of court school finance cases. Nonetheless, the main objective of the chapter is describing both the historical and contemporary strands of relevant research into funding and student outcomes. We provide a systematic review of the contemporary quantitative analyses linking resources to outcomes. We then provide an outline of some open questions along with our conclusions about the results of the existing high-quality analyses of the school resource questions.

2 Attendance, Finance, and Outcomes

In order to frame the school finance discussion, we begin with a brief description of the nature of financing of schools in the United States. The overall picture of enrollments, structure of the schools,

and funding shows some significant changes over time. But the aggregate picture also hides an enormous heterogeneity across the states. Because of the central role of states in setting policy and in funding of the schools, this heterogeneity provides an important backdrop both for the analysis of school finance issues and for decision making in the schools.

2.1 The Shape of U.S. Schooling

Public school enrollment in the U.S., while rising during the 1990s, reached 50 million students in 2013 and stabilized there until the COVID-19 pandemic hit in 2020. The full extent of reaction to the pandemic is not yet known, but public school enrollment fell by three percent from Fall 2020 to Fall 2021 and remained at the lower level through Fall 2022.

Students are spread very unevenly across states and, within states, across separate local school districts. At the state level, Vermont had a total of 82,000 students while California had six million. The prime operating level is the school district of which there were 13,452 in 2019, down from 117,408 in 1940. Moreover, the states are broken up into widely varying numbers of local districts. While Hawaii and the District of Columbia each have only one school district, five states had more than one thousand districts.

But even these aggregate variations understate the degree of heterogeneity in the schools. The growing importance of school choice leads to even more decentralized operation of education. The public school district is the prime operating unit, but it does not cover the full provision of educational services. First, beginning in 1991, charter schools were established in Minnesota, and the model spread across the country. Charter schools are public schools that operate with varying degrees of autonomy, depending on the state. Typically, charter schools are free to operate outside of many of the education regulations in a state and importantly can, independent of local teacher unions, set their own requirements for teacher preparation, their own salary schedules, and their personnel rules. They receive public funding, and they are almost always required to take all applying students or to randomize admissions if more students apply than they can accommodate. They are required to participate in the state student assessment systems.

In addition to the charter schools, students can attend private schools or be home-schooled. While changing, private schools almost always receive no direct public funding, as is the case for home-schooling. These parts of the system are generally very unregulated, and they can set their own curricula and standards. They generally do not participate in state student assessment systems.

Figure 1 shows the substantial changes in the structure of U.S. schools in the 21st Century in terms of parental choices that interact with school finance.[1] There has been a steady rise in charter school attendance with relatively stable home-school attendance and declines in private schooling. The

[1] There are more dimensions of choice, but they do not interact significantly with overall financing. Most importantly, while districts with assigned attendance zones for neighborhood schools predominate, many districts have magnet schools that draw students from the entire district to attend schools with a specialized focus or have open enrollment across all schools in the district (see Abdulkadiroğlu and Andersson (2023)). Such choices in general do not affect the total funding for the district whereas the choices in Figure 1 will affect funding for traditional districts. They do have impacts on school performance; see Angrist, Hull, and Walters (2023), CREDO (forthcoming 2023).

private school attendance is one-quarter nonsectarian and three-quarters religious based with the religious component evenly split between Catholic and other denominations.

Note, however, that these data are all pre-pandemic. With the pandemic, traditional public school attendance fell while the other choice options increased. Within the public school sector there was also a shift from the traditional public schools to charter schools. The long run distribution is yet unclear.

2.2 Revenues for U.S. Education

The structure of the educational sector and the attendance patterns that were highlighted relate directly to school finances. Because private schools and homeschooling are not publicly supported (to any significant degree), any increased attendance in these sectors relieves state and local governments of resource demands, although it may also reduce political support for public schools.[2]

Figure 2 traces revenues for the public schools from 1960-2019. The bulk of funding comes from state and local revenues which each correspond to roughly 45 percent of per pupil funding. The federal share, which began rising in the 1960s as the federal government assumed a larger role in financing schools for disadvantaged students and subsequently for special education students, rose around the 2008 recession and then returned to their historic levels. While not shown, the federal government also contributed large additional amounts of temporary funds with the onset of the pandemic in 2020. The steady increase in per pupil funding over the entire period puts public school revenues per student in 2019 at over four times that in 1960 in real terms. In fact, except for the dip in school revenues after the end of federal support for the 2008 recession, real per pupil spending has risen continuously for over 100 years. State revenues come from a variety of sources that differ across the fiscal structures of the different states. At the same time, with few exceptions, property taxes remain the dominant source of local revenues. Public school spending incorporates both traditional public schools and charter schools. For a variety of political and institutional reasons, charter school spending is systematically below that for traditional public schools, although there is debate about the exact magnitude of differences.[3]

The aggregate data hide the wide variation that is seen at the state level. States differ significantly in how revenues are raised and in the level of spending. Table 1 shows the extent of

[2] The sharp funding distinction between public and private schools has been breaking down as various states have introduced funding for special needs students attending private schools, limited forms of school vouchers, and more general support of students seeking education outside of traditional public schools. Importantly, a growing number of states have established education savings accounts that provide funding that can be used at private schools for some subset of students or for all students in a state; see https://www.edchoice.org/school-choice/types-of-school-choice/education-savings-account/ [accessed February 10, 2023].

[3] Available research indicates that charter schools receive considerably less funding per-pupil than traditional public schools in most states, though levels of funding vary widely across charter schools (Miron and Urschel, 2010; Nelson, Muir, and Drown, 2003; Belfield, 2008). Other researchers argue that these assessments ignore the large role that private philanthropic contributions play in funding charter schools in some regions (Baker, Libby and Wiley, 2015; Baker and Ferris, 2011). These researchers also note that the demographic makeup of charter school students differs from that of traditional public schools in that charter schools typically serve populations with fewer economically disadvantaged or special education students (Baker and Ferris, 2011), meaning that funding needs may be lower (Baker, Libby, and Wiley, 2015; Baker and Ferris, 2011). For analysis reaching the opposite conclusion, see Dills et al. (2021).

compositional differences in school funding. Typically, most of the revenue is derived from state and local sources with the federal government contributing a smaller portion but the federal share across states differs from 4 to 15 percent of funding. States like Hawaii with its one district and Vermont provide almost all funding at the state level, while funding for schools in Washington, D.C. is provided almost entirely at the local level. 15 percent of the funding for Alaskan schools comes from the federal government, the highest percentage of all states. Figure 3 illustrates the distribution of state per-pupil spending levels in the 2018-19 academic year. Northeastern states spend over $15,000 per student, significantly higher than the $9,000 to $11,000 per pupil spent by the majority of southern states.

2.3 Student Performance

The United States has a long tradition of assessing student performance. The National Assessment of Educational Progress (NAEP) is known as the Nation's Report Card. Going back to 1978, the Long Term Trend (LTT) assessment of NAEP made it possible to get representative national data for math and reading performance of students aged 9, 13, and 17. Beginning in 1992, a second version of NAEP, called Main NAEP, was started with testing of math and reading in grades 4 and 8.[4]

Table 2 provides data on NAEP testing results both in terms of changes in standard deviations (SD) and in terms of these changes relative to school expenditure. The pre-pandemic results fall into two distinct clusters. There are strong gains in the level of math performance for younger students – age 9 (grade 4) and to a lesser extent age 13 (grade 8).[5] But there are much more modest gains for reading of all ages and for age 17 math. The scores cover different periods of time, so it is also useful within this discussion to place them in comparison to the spending on schools. When normalized by spending over the relevant time periods, the younger cohort math gains are all greater than 0.07 SD per 10 percent larger spending, while the remaining gains are all less than 0.03 SD per 10 percent larger spending.

The results were, unsurprisingly, dramatically altered by the COVID-19 pandemic. The Main NAEP had testing in Spring 2019 (included in Table 2) and Spring 2022. In math and reading for both grade 4 and grade 8, average scores fell dramatically with the largest declines being recorded for math performance (Table 3). Grade 8 (grade 4) gains from 1990 through 2022 were down to 0.33 SD (0.72 SD). For reading, virtually all gains since 1992 were erased by the pandemic; the 1992-2022 gain was 0.01 SD for grade 8 and 0.02 for grade 4. It is of course difficult to know how to interpret the scores after the pandemic, but they do suggest that the added funds over the pandemic period were insufficient to overcome the learning disadvantages of the pandemic period.

The achievement gains in Table 2 are unconditional changes in student performance. In interpreting these performance data, however, it is important to note that achievement is a function not only of schools but also of parents, peers, and neighborhoods. Thus, while the scores normalized by spending give a benchmark about how spending and performance have moved together, they obviously

[4] Main NAEP has much larger samples of students in order to provide state-by-state performance data. It has also tested 12th grade reading and math and various other subjects such as history, civics, and geography on a less regular basis and using significantly smaller samples of students. These additional tests do not provide consistent time series data.

[5] LTT NAEP is age based while Main NAEP is grade based.

do not provide information about the causal impact of spending. That is the subject of the subsequent sections.

The national achievement data mask the fact that there are dramatic differences in achievement across states. Figure 4 arrays the eighth grade math performance on the NAEP tests for each state in 2022. The differences in performance across states is very large. By conventional estimates, the difference in performance between Massachusetts (the top performing state) and New Mexico (the bottom performing state) translates in 2-2.5 years of education at the eighth grade.[6]

One related pattern that does take into account some of non-school factors is the historical evolution of achievement gaps by socio-economic status (SES). Concerns have been raised that the widening of the U.S. income distribution led to expanding SES-achievement gaps (Reardon (2011)). That concern, however, appears unfounded as test information that is linked over time shows a slow shrinking of gaps for birth cohorts born between 1961 and 2001 (Hanushek et al. (2022b)).

3 Special Role of Courts

The United States stands alone in the role that courts have played in school policy decision making. The power of the courts to intervene is derived solely from their authority to enforce certain rights under both federal and state constitutions, such as the right to equal protection of the laws. We review the history of school finance court cases both because the courts remain a continual force in finance decisions and because their existence plays into some current analytical strategies investigating the impact of finance policies on student outcomes.

3.1 Federal Courts

The federal courts have not had a consistent long-run impact on school finance. At the federal level, the relevant judicial actions can be traced to the landmark 1954 desegregation decision by the U.S. Supreme Court, *Brown v. Board of Education.*[7] That court case applied the Fourteenth Amendment (equal protection clause) of the U.S. Constitution to the *de jure* segregation of schools. Initially, the desegregation decrees of the court were directed at reassigning students to eliminate one-race schools, integrating faculty and staff, and ensuring equal allocation of facilities and other resources. But beginning in the early 1970s, the federal courts also began to address funding issues and to order states and local school districts, as part of their desegregation plans, to improve the quality of education offered in predominantly black schools by providing extra funding for "educational enhancements." Thus, the federal courts initially entered into school finance decisions through desegregation orders under the equal protection clause of the U.S. Constitution. The spending requirements for desegregation purposes reached an extreme in Kansas City during the 1990s. As a result of its extraordinary court-ordered funding, the Kansas City Municipal School District went from spending at the national average to spending more per pupil than any of the 280 largest school districts in the country. But the federal courts subsequently moved away from such rulings. In 1995, a ruling in the

[6] The rule of thumb, derived from scores on vertically-aligned tests, is the one standard deviation of achievement is equivalent to 3-4 years of school.

[7] A review of the history of school finance court cases can be found in Hanushek and Lindseth (2009). This discussion provides a synopsis of the more complete analysis in that book.

case of *Missouri v. Jenkins* by the U.S. Supreme Court ended this spending that the State of Missouri had been required to provide, and more generally the funding decisions related to desegregation receded.

The more general issues of school finance outside of desegregation considerations were brought into federal courts in 1968. The Texas system of funding schools through local property taxes was challenged in federal court as discriminatory and in violation of the equal protection clause of the Fourteenth Amendment. In 1973, the United States Supreme Court rejected that claim in *Rodriguez v. San Antonio,* ruling that school funding did not concern a fundamental right under the federal constitution that does not mention education in its text. Therefore, education was ruled to be appropriately a matter left to the states.

3.2 State Courts

At the time of the federal court decision in *Rodriguez*, civil rights groups and property-poor school districts had already begun to pursue their equal protection claims in state courts under state constitutional provisions. The claims pursued in the state courts argued that the education funding "pie" should be divided more equally among a state's school districts and (in the language of Coons, Clune, and Sugarman (1970)) rested on the premise that the quality of a child's education should not depend upon the wealth of one's neighbors.

The earliest of these state court "equity" cases was *Serrano v. Priest,* in which plaintiffs in 1968 challenged California's education funding system. In California, as in Texas, the public schools were financed largely through a combination of local property taxes and state revenues. Most funding came from local property taxes, which varied greatly from school district to school district, depending largely on differences in the property tax base.[8] While California employed a foundation formula with student-weighted state funding designed to moderate disparities in property tax bases, the compensation for differing tax bases was relatively low, leading to wide variation in local revenues.[9]

Similar equity court cases were pursued in almost all states, meeting various degrees of success (see below). Ultimately, plaintiffs were successful in less than half of these cases. The equity lawsuits that found a state's financing to be unconstitutional did not, however, always lead to increases in school funding. First, most funding disparities were not primarily driven by the state funding but instead were the result of a subset of districts raising additional money to support their local schools. Thus, if a court ordered more equality in spending across districts, it could be achieved by limiting spending of districts with the largest revenues, leaving poorer districts unchanged. If the results of equity suits did not

[8] The property tax base combines the value of residential properties with the value of commercial and industrial properties. Therefore, poor districts are not the same as poor people, but depend in part on the distribution of nonresidential property. Most states compensate partially for differences in the property tax base. Three basic funding mechanisms characterize the options – and most states use a combination of them. Categorical aid provides funds for districts based upon specific identified needs; foundation aid compensates for differing tax capacity of the local district; and variable matching aid adjusts state support for both differing tax capacity and for the taxing decisions of the local district. Foundation plans account for the bulk of state funding for local districts. See Hanushek (2002) for a discussion of the alternative financing approaches.

[9] For a discussion of the use of property taxes, see Fischel (2006). See also the discussion about the relationship between equalization suits and referenda to limit school spending (Fischel (2006) along with Fischel (1989), Silva and Sonstelie (1995)).

expand the pie, there would be both winners and losers. Second, the definition of equity was unclear since horizontal equity might still call for more spending for districts with greater at-risk student populations including those with special needs, English language challenges, and the like.

These arguments supported a different kind of court case around the concept of "adequacy." These suits were fundamentally different than the state court "equity" cases that preceded them. They had their genesis not in the equal protection clause of state constitutions, but in the "education clause" of state constitutions. Virtually every state constitution requires that the state or its legislature provide some form of free public education for the children of the state. This requirement is normally couched in very general terms, such as the requirement that the state or legislature provide a system of "free common schools" (New York), "cherish the interests of literature and the sciences" (Massachusetts), "make suitable provision for finance of the educational interests of the state" (Kansas), or establish "a complete and uniform" and "thorough and efficient system" of public schools (Wyoming).

In adequacy cases, the courts are called on to decide what level of education is required under the vaguely worded state constitutions, whether the state provides such an education, and, if not, what needs to be done to remedy the situation. In other words, the school finance formula might provide equal education resources across districts (including after adjustments for demographic differences across districts), but these resources might be deemed insufficient to meet the requirements of the education clause of the state constitution.

The following section describes the range of state school finance cases that have been decided through the end of 2021. It provides information about both the types of cases and their disposition.

3.3 Quantitative History of Court Cases

Following the entry into school finance decisions through 2022, state courts have been involved in 205 identifiable litigations.[10] These cases have all been brought under the individual state constitutions. The pattern of court cases over time along with the decisions is provided in Figure 5. Cases are identified by the year in which the final decision was made.[11] Note, however, that cases were initiated at varying times before the final decision that, on average, comes 3.5 years after the case was first launched.

There has clearly been an increase in cases over time. While the decades of the 1970s and 1980s had less than 20 cases per decade, the numbers grew to over 50 per decade in the 21st Century.

Also evident is a slightly lower rate of success for the plaintiffs (47 percent) over the entire period. While there is no clear time trend in decisions, the verdicts over the last two decades tend to favor defendants.

[10] Cases have been coded based on decisions found in standard legal references. For the most part, the plaintiffs are interested parties who sue the state to change the existing school finance policies. The defendants are generally representatives of the state executive and/or legislative branches. See Hanushek and Wirtz (forthcoming) for an extensive discussion of school finance court cases.

[11] Data in the chart represent 198 cases decided by the courts. Seven cases either reached a settlement or ended because of a separate legislative action.

The cases have not been evenly distributed across the country. California, New Jersey, New York, and Kansas have each had 10 or more separate cases, while 16 states have had two or fewer cases. As is apparent in Figure 6, there is no obvious regional pattern across the states in the number of cases.

As noted, the nature of the litigation shifted over time with the early cases being pure equity cases and the later cases being adequacy-based or a combination of equity and adequacy. Table 4 summarizes both the type of court case and whether the latest decision was for the plaintiffs or for the defendants.[12] Across all of the state court decisions, 53 percent were decided for the defendants, which in general implies retaining the system of finance in place at the time of the decision. For the pure equity decisions, 59 percent ultimately favored retention of the current system. But, those cases combining both equity and adequacy yielded a majority of decisions for the plaintiffs.

3.4 Interactions with finance

It is informative to see how the court cases interact with school spending. While the equity cases have their basis in the distribution of spending within states, the adequacy court cases focus on the level of funding, and whether, in the opinion of the court, the funding is adequate to meet the educational goals of the state constitution. To learn whether base levels of funding impact levels of litigation, we compare the distribution of court cases from states spending below the national average at the filing date to those spending above the national average. As seen in Table 5, a greater portion of the adequacy or combined adequacy and equity cases are launched in states that spend below the national average.[13] This difference would be consistent with court cases being more common in places with greater needs. Adequacy cases are noticeably less likely to be decided for the defendant (retaining status quo) in the high spending states. This disparity in findings by spending levels suggests that it is not just pure resource issues driving the court decisions.

The courts have been very active in school finance, but it is important to keep in mind exactly where they enter into policy discussions. Their role has throughout their history focused on the level and distribution of funds. This put the focus solely on bolstering and equalizing inputs, not on maximizing outcomes per se. Yet, a central element of much of the litigation has been discussion of how overall funding affects student outcomes. The following sections address this fundamental issue.

4 Resources and Outcomes (Historical Studies)

Education finance policy discussions were radically changed with the publication of *Equality of Education Opportunity* (Coleman et al. (1966)). This seminal government report, commonly called the "Coleman Report" after its primary author James Coleman, introduced the idea that understanding inequities in education should come from consideration of student outcomes. But the aspect of the report that received the most public and scholarly attention was its controversial finding that schools (as measured by various resources) had little impact on student achievement. Instead, student

[12] Some current cases are under appeal, and the decision refers to the last decision as of September 2022. Seven cases are not included because they did not have a final decision owing to a settlement or legislative action that ended the case. In general, the plaintiffs have brought suit to change the funding formula while the defendants represent the state government acting to stop the suit and to retain the current funding system.

[13] The number of cases in the discussion of spending patterns differs from the total number of cases filed because of the lack of relevant spending data for the five years preceding the filing in some cases.

achievement was most importantly related to family background and, to a lesser extent, peers in the schools.

The Coleman Report was a federal government study mandated by the Civil Rights Act of 1964 which stated:

> SEC. 402. The Commissioner shall conduct a survey and make a report to the President and the Congress, within two years of the enactment of this title, concerning the lack of availability of equal educational opportunities for individuals by reason of race, color, religion, or national origin in public educational institutions at all levels in the United States, its territories and possessions, and the District of Columbia.

The U.S. Congress intended for the U.S. Office of Education (the predecessor of the Department of Education) to record differences in school facilities and school personnel for students of different races and backgrounds in the previously segregated schools of the U.S. South. The resulting report was quite different from any prior education reports. It developed surveys for children, parents, and school administrators. Importantly, it also introduced a battery of achievement and ability tests that were administered to some 600,000 students spread across the country in grades 1, 3, 6, 9, and 12. Moreover, it did not stop at this point but instead proceeded to estimate statistical models of how survey items were related to achievement.

Its initial efforts today appear quite primitive and obviously flawed, but at the time they were revolutionary. The Coleman Report moved attention to student outcomes instead of simply looking at school inputs as a measure of school quality. It also directly identified a multitude of factors affecting student achievement including families and peers. The full impact of these facets of the Coleman Report were not completely understood for some time after its publication. While heavily criticized on methodological grounds, it is still heavily cited for its findings a half-century after its publication.[14]

The results of the Coleman Report, frequently interpreted as suggesting that "schools do not matter," led to a large volume of related work. These follow-on studies were, perhaps naively, directed at understanding basic elements of schools that would lead to better student outcomes and were largely motivated by ideas of improved decision making and policies. Because systematically testing the achievement of students was not commonplace until the last decade of the twentieth century, the follow-on research to the Coleman Report included many samples of convenience constructed with limited data relevant for very specific circumstances.[15] And even by then-prevailing standards, the empirical analyses were of highly variable quality.

While these studies emphasized varying aspects of education, the overarching theme of them was estimation of an educational production function. In particular, these studies quite uniformly recognized that education was not just something that occurred in schools. As emphasized by the Coleman Report, families were very important in education. But the main research focus was the school with an attempt to understand how different components of schools and their resources affected student outcomes.

[14] For early critiques of the methodology, see Bowles and Levin (1968), Cain and Watts (1970), and Hanushek and Kain (1972).

[15] See Hanushek (1997) for a description of the different samples and facets of this research.

The general framework of analysis of educational performance considers a general production function such as:

$$A_{it} = f_{\rho}(F_i^{(t)}, P_i^{(t)}, S_i^{(t)}, A_i) + \nu_{it} \quad (1)$$

where A_{it} =performance of student i at time t, $F_i^{(t)}$ =family inputs cumulative to time t, $P_i^{(t)}$ =cumulative peer inputs, $S_i^{(t)}$ =cumulative school inputs, A_i =innate ability, and a stochastic term, ν_{it}.[16] Importantly, because policies differ significantly across states, the precise relationship of inputs to student performance is modelled as depending on the policy environment (ρ) of the schooling that determines how the various inputs enter into the outcome of the process ($f_{\rho}(\cdot)$).

A key initial issue is how student performance is measured. A prime justification for the attention to education is its hypothesized effects on labor market outcomes. The question remains about how best to measure educational output for understanding production relationships and policy options. Most of the historical analysis has not been related to subsequent earnings or labor market experiences.[17] Instead, the analyses have focused on test scores, school completion, or other intermediate outcomes, with test score analysis being the most common. This focus on intermediate outcomes, however, does not seem too problematic because skills measured by test scores have been shown in a range of studies to be closely related subsequent economic outcomes (see Hanushek and Woessmann (2015) on economic growth and Hanushek, Schwerdt, Wiederhold, and Woessmann (2015, 2017) on individual labor market earnings).

This general production function structure motivated an extensive series of empirical studies. The typical empirical study collected information about student performance and about various measured educational inputs and then attempted to estimate the characteristics of the production function using econometric techniques. The immediate wave of academic studies produced in reaction to the Coleman Report included studies of highly-varying quality.

Three aspects of this formulation are important to emphasize. First, a variety of influences outside of schools enter into the production of achievement. Second, the production process for achievement is cumulative, building on a series of inputs over time. Third, the policy environment might affect how resources are converted into student outcomes. Each of these is important in reviewing the various specifications and interpretations of analyses educational production functions.

If we take Eq. 1 as the appropriate underlying model, the fundamental concern can be simply stated as general problems of omitted variables that imply a correlation of ν_{it} with the included inputs and, most importantly, with the measures of schools, $S_i^{(t)}$. This potential problem creates varying interpretative issues for the historical estimates of the impacts of varying school factors, where the

[16] See the more general discussion in Hanushek (1997) and the related but somewhat different formulation in Todd and Wolpin (2003).

[17] Exceptions are found, but these studies are generally aggregated to high levels such as the state level, and, as discussed below, this introduces a wider set of analytical concerns.

severity of the problems can be readily seen in many instances by the analytical structure of the analysis.

The cumulative nature of achievement is particularly important. Because the learning in any time period builds on prior learning, analysis must take into account the time path of inputs. This places heavy demands on measurement and data collection, because complete and accurate historical information is frequently impossible to obtain. A large portion of historical production function studies ignored the cumulative input issue and analyzed purely cross-sectional achievement differences.

The cumulative nature of the production process has been a prime motivation for considering a value-added formulation. At least in a linear version of Eq. 1, it is possible to look at the growth in contemporaneous performance over some period of time, instead of the level of performance. This growth can then be related to the flow of specific inputs. A simplified view of the general value-added formulation can be written as:

$$A_{it} - A_{it^*} = f_P(F_i^{(t-t^*)}, P_i^{(t-t^*)}, S_i^{(t-t^*)}) + (\nu_{it} - \nu_{it^*}) \quad (2)$$

where outcome changes over the period (t-t*) are related to the inputs applied over the same period (e.g., $S_i^{(t-t^*)}$). Note that this formulation dramatically lessens the data requirements. It also eliminates anything that appears as a fixed effect in the level of achievement (eq. 1), something that is very important given the often limited measures of family inputs that are available.[18]

Alternative formulations estimate models with prior achievement, A_{it^*}, on the right-hand side and allow for a coefficient on lagged achievement that is different than one (Hanushek (1979)). This latter approach has the advantages of allowing for different scales of measurement in achievement during different years and of introducing the possibility that growth in performance differs by starting point. It has the disadvantages of introducing measurement error on the right hand side and of complicating the error structure, particularly in models relying on more than a single year of an individual's achievement growth.

This general production function structure corresponds to the majority of analyses of the impact of resources on student outcomes that were conducted in the 20th Century. It is useful to consider the results of this estimation and to evaluate what can be concluded from these about resource policies.

4.1 Summary of Historical Research

The analysis of school resources began to appear in publications soon after the Coleman Report. We provide a high level summary here because these studies represent the received wisdom that has entered into policy discussions and the related court and legislative proceedings. This research summary also facilitates comparisons to more recent studies of resource effects.

[18] This formulation presumes that innate abilities are constant and thus fall out of achievement growth. With more information on variations over time, it is also possible to allow for ability differences in growth (Rivkin, Hanushek, and Kain (2005)).

This research peaked in the late 1980s and early 1990s at a time when the results provided a consistent statistical picture, and there were few incentives for individuals or journals to add additional analyses. Hanushek (2003) provides the most complete picture of the range of historical studies. The thrust of these early studies was investigation of key parameters related to school resources. The dwindling numbers of relevant studies of this genre after the mid-1990s did not led to a different picture.

Estimates of key production function parameters can be found in an exhaustive search of 376 separate published estimates, found in 89 separate articles or books up to 1994 (Hanushek (2003)).[19] The estimated relationships differ in a variety of substantive ways (by measure of student performance, by grade, by included measures of resources). These studies also vary widely in quality, as generally captured by methodology and adequacy of data. [20]

Table 6 presents summary of the overall results of historical estimation of educational production functions. For expositional purposes, parameters are divided into: 1. Real classroom resources (teacher-pupil ratio, teacher education, and teacher experience);[21] 2. Financial aggregates (teacher salary and expenditure per pupil); and, 3. Other (facilities and administration). This breakdown also facilitates comparisons to the more recent genre of school resource studies that follows in the next section. This table summarizes the extant studies by dividing all of the relevant parameter estimates into those that have a statistically significant positive effect (the expected sign for each if more resources are beneficial), statistically significant negative effect, and statistically insignificant.[22]

In terms of real classroom resources, only 9 percent of the estimates considering the level of teachers' education and 14 percent of the estimates investigating teacher-pupil ratios find positive and statistically significant effects on student performance. These relatively small numbers of statistically

[19] A subsequent controversy centered on how to summarize the results of studies. Krueger (2000) introduced a different measure of study quality. His proposed measure was the number of separate parameter estimates in a given published analysis. So, for example, a publication that included estimates from a production function for eighth grade reading and one for high school graduation would necessarily be lower quality than a publication that only reported on third grade mathematics. Direct comparisons of the estimates based on the alternative weighting of the two approaches, nonetheless, indicates that they give generally similar results except for the heavy weighting of the low quality studies of state level expenditures (see below).

[20] A more complete description of the underlying studies can be found in Hanushek (1997).

[21] The real classroom resources provide direct information about spending at the classroom level since teacher salaries are systematically related to teacher education and experience and the teacher-pupil ratio indicates the number of teachers required for a given number of students. In general, spending per se is never directly computed at the classroom or even school level. Additionally, as discussed below, the estimates of the impact of spending per pupil at the district or state level tend to be lower quality studies, implying that a focus on the real school and classroom measures provides a better way to understand the role of resources. Over time, increases in teacher-pupil ratios have been the largest component of increases in expenditure per pupil (Hanushek and Rivkin (1997)).

[22] Various forms of meta-analysis look at different ways to aggregate the results including both statistical significance and quantitative effects. Hedges, Laine, and Greenwald (1994) argue that the approach here, sometimes labeled "vote-counting," can be misleading because it can be prone to Type II errors (accepting a null hypothesis that is not true). This issue is particularly salient when the true underlying parameter is small. We return to these issues below when we discuss more relevant subsets of estimates.

significant positive results are balanced by another set finding statistically significant negative results—reaching 14 percent in the case of teacher-pupil ratios.[23]

A higher proportion of estimated effects of teacher experience are positive and statistically significant: 29 percent. The statistically significant estimates of experience generally reflect a consistent finding that teachers improve in their first few years of teaching but that afterwards there is no clear improvement. Importantly, 71 percent still indicate either worsening performance with experience or less confidence in any positive effect. In sum, the vast number of estimated real resource effects in these historical studies gives little confidence that just adding more of any of the specific resources to schools will lead to a boost in student achievement.

The financial aggregates provide a similar picture. There is very weak support for the notion that simply providing higher teacher salaries or greater overall spending will lead to improved student performance. Per pupil expenditure has received the most attention, but only 27 percent of the estimated coefficients are positive and statistically significant. In fact, 7 percent even suggest some confidence in the fact that spending more would harm student achievement. As discussed below, analyses involving per pupil expenditure tend to be the lowest quality of these historical studies, and there is substantial reason to believe that even these results overstate the true effect of added expenditure. The relatively small number of estimates of the effect of facilities or administrative inputs come from very idiosyncratic measures of these inputs. As a whole, they provide little support for having a strong influence on student outcomes.

These overall estimates should nevertheless not be over-interpreted. A large proportion of them come from simplistic cross-sectional estimates following Equation 1. As such, they are very prone to omitted variable biases. (Note, however, that if the omitted factors tend to be positively correlated with the included resources, the estimates would tend to be biased upwards and thus actually to overstate the true effects of the identified resources).

For purposes of comparing these estimates with the more recent research on spending, it is useful to go further into the results of these early studies. Estimates of the impact of expenditures per pupil as found in Table 6 do not come from the classroom or even the school level because spending data are not measured at those levels. Instead they come from estimates employing data aggregated to district or state level. Moreover, the analyses are heavily weighted toward analyses across states. Clearly, the policy environments across states, i.e., $f_{\rho}(\cdot)$ in Eq. 1, differ dramatically, and different policy environments may be correlated with the resource measures in the estimation.

It is possible to get some sense of the importance of omitted factors – particularly as related to the policy environment – by looking in more detail at the teacher-pupil ratio and the expenditure per pupil estimates. Specifically, Hanushek, Rivkin, and Taylor (1996) demonstrate that aggregation alters the degree of omitted variables bias, even when the true marginal impacts of included variables are constant across different levels of aggregation. Omitted variables have their strongest effects on estimates when the data are aggregated to the level of the omitted factors (such as when the data are

[23] While a large portion of the studies merely note that the estimated coefficient is statistically insignificant without giving the direction of the estimated effect, those statistically insignificant studies reporting the sign of estimated coefficients are split fairly evenly between positive and negative.

aggregated to the state level and state-level determinants of students' performance are neglected). In this case, aggregation increases the magnitude of omitted variable biases. Given the dominant role of states in school organization, financing, and regulation, it is likely that state-level resources are correlated with a variety of important state influences on school performance. Therefore, aggregation-induced changes in the magnitude of omitted variables bias provide a plausible explanation for the pattern of school resource results.

Table 7 breaks down the estimates of teacher-pupil ratio and expenditure per pupil by level of aggregation of the data (state level or less than state level) and by whether the estimation samples come from multiple states (i.e., multiple policy environments). What is immediately obvious is that estimates of either measure of school resources are much more likely to be positive and statistically significant for models that use cross-state samples and that measure the resources at the state level. These findings are consistent with state policies being very important for the effectiveness of school resources, and, when not considered, causing substantial bias in the estimates. Again, the collection of these early results, while potentially offering some information about the impacts of various school factors, is prone to potentially serious identification problems.

4.2. Value-added Estimates of Production Parameters

There is, however, a subset of the historical education production function estimates that provides more reliable information about the impact of the real resources. A number of analyses have pursued the value-added formulation of Eq. 2. This analytical approach directly deals with historical inputs and eliminates any fixed inputs such as overall family effects. By focusing on the impact of flows of school resources, it provides clearer evidence on the impact of resources on student outcomes.

Table 8 summarizes the estimated impacts of real resources (class size, teacher experience, and teacher education) on achievement as identified in value-added models estimated across individual students and individual classrooms. The top panel includes all of the estimates found in these studies, while the bottom panel is restricted to estimates from individual states – thus, eliminating the possible influence of differential policy environments, $f_\rho(\cdot)$.

For these high quality estimates, there is again an indication that initial years of teacher experience are valuable in terms of student achievement, but none of the other standard school resources show any consistent relationship to achievement. In the lower panel that has within-state studies (i.e., within-policy environment studies), there is even a slightly stronger indication that smaller class sizes are harmful.[24]

These results lead to three major conclusions. First, taken as a whole, the entire set of education production function estimates is very difficult to interpret. The likelihood of biased estimates based on partially specified cross-sectional models is clear, and it is very difficult to ascertain the degree

[24] Note that in general teacher-pupil ratios are not the same as class size because teachers can be assigned to non-classroom activities and, where there is subject-specific teaching, may not teach as many sections of students as the total number of sections that students take. In the case of estimates for individual classrooms, however, teacher-pupil ratio indicates class size.

of bias in the estimates. Second, the historical estimates do not provide reliable estimates of the impact of differential expenditure per pupil, because the evidence is heavily weighted toward the lowest quality estimates. But these aggregate estimates do provide a clear indication of the underlying importance of correlated policy environment factors at the state level. Third, there is little evidence of consistent impacts of added resources (as commonly measured).

Ignoring issues of methodological problems, the variation in impacts of resources across studies could arise because of differences in the associated regulations, policies, and organization of the sampled schools; because of different choices by sampled schools and districts in the use their funds; or because of an interaction between these. In particular, among other things, federal, state, and local regulations, the supply of school personnel, local union contracts, and community and parental preferences place constraints on the set of resource allocations that are possible. Within this institutional environment – which will differ across states, districts, and schools – local decision makers of differing ability are making resource allocations and management decisions. The sampled school experiences that underlie the various results presented previously undoubtedly embody a wide range of environmental and personnel constraints that enable or limit the effective use of any added resources. We refer to these possible effectiveness-related mechanisms that enter into the pattern of research results collectively as "how" the resources are used. From the variation in estimated impacts of the real resources in the historical value-added studies we might infer that how resources are used is of primary importance. Importantly, there is little historical indication of the best ways or even worthwhile ways of ensuring productive usage.

These issues become clearer in the next section about contemporaneous studies. The historical studies as a group are plagued by methodological issues, making it difficult to distinguish between underlying bias in the estimates and genuine variations in effectiveness as determined by how the resources are used. The more recent studies focus on minimizing any bias in estimation of specific resource impacts, thus leading to a more direct consideration of the importance of how resources are used.

4.3 Teacher Effectiveness

At this point, the evidence appears to be generally consistent with the Coleman Report conclusion that differences in schools are not important, but such a conclusion would be a misinterpretation of the evidence. The historic evidence is consistent with a finding that measured resource differences are not closely related to student outcomes, but that is different from saying that schools do not matter. This section provides evidence on differences in teacher effectiveness and makes the point that, contrary to the Coleman Report conclusion, schools are very important.

An expanding line of research considers differences in teacher effectiveness. It is important because it identifies key elements of the impact of schools on student outcomes. This research indicates that schools can have substantial impacts on student learning, but at the same time the impact of schools is not well-characterized by differences in the measured background, characteristics, and experiences of teachers.

The general formulation of this line of research can be written as an extension of the value-added version of the educational production function:

$$A_{it} = f_{\rho}(A_{it^*}, F_i^{(t-t^*)}, P_i^{(t-t^*)}, S_i^{(t-t^*)}) + \delta_{\tau} + \nu_{it} \quad (3)$$

In Eq. 3, δ_{τ} is a fixed effect for teacher τ, and the prior achievement (A_{it^*}) is written on the right hand side (as found in most of actual estimation).

The earliest work on teacher value-add exploited specialized data sets (e.g., Hanushek (1971), Murnane (1975), Armor et al. (1976)). The early studies had relatively small and unique samples and covered different districts and regions, yet they had remarkably similar findings about the distribution of teacher effectiveness as estimated by Eq. 3 (Hanushek and Rivkin (2010)).

Research into teacher value-added modeling expanded dramatically over the first two decades of the 21st century. This expansion partly reflected the significantly increased availability of school and state administrative data that tracked the performance of individual students. Such administrative data were routinely produced because of the requirements for school accountability under the federal mandates of No Child Left Behind.[25] These data facilitated extensive investigations of teacher value-added in instances where students were linked to their teachers.

The expansion of this research has gone in two basic directions. One line of research has focused simply on the interpretation of variations in estimated value-added and has focused on the stability and unbiasedness of results for individual teachers. This focus recognized the fact that many states began to call for the evaluation of individual teacher to be based at least in part on estimates of teacher value-added. The second line of inquiry focused on what factors could explain variations in teacher effectiveness. This latter line of research was motivated by various recommendations for the training, certification, and pay of teachers and in particular whether current practices in these areas were consistent with observed variations in value-added.

The investigation along both of these lines of research has been extensive and has been thoroughly reviewed in multiple publications (see, for example, Staiger and Rockoff (2010), Harris (2011), Hanushek and Rivkin (2012), Jackson, Rockoff, and Staiger (2014), Koedel, Mihaly, and Rockoff (2015), Bacher-Hicks, Chin, Kane, and Staiger (2017), and Bacher-Hicks and Koedel (2023)). The key conclusions of these overall evaluations are important both in the context of educational production functions and of their relevance for school finance policy in general.

First, there is no doubt that that teachers differ widely in their effectiveness as measured by gains in student achievement. Second, teachers generally become more effective in their first few years of teaching, but changes in effectiveness quickly plateau with additional experience. Third, except for the impact of initial experience, few measures of background (e.g., certification, advanced degrees,

[25] The No Child Left Behind Act of 2001 required regular annual testing of students in grades 3-8. These data were collected by states and gradually became available to researchers through different access regulations that were often dictated by privacy considerations. Nonetheless, not all states linked student performance to individual teachers so, even when administrative data were available, they did not always support estimation of teacher value-added.

amount of professional development, or salary) are significantly related to differences in teacher value-added. Fourth, variations in teacher effectiveness are the largest component of overall school quality.

To understand the magnitude of teacher differences, estimates of teacher effectiveness have been put in terms of potential impacts on future labor market outcomes for students (Gordon, Kane, and Staiger (2006), Goldhaber (2009), Hanushek (2011), Goldhaber and Hansen (2013), and Chetty, Friedman, and Rockoff (2014)). While these estimates are all simulations of one sort or another, they reach similar conclusions: Replacing the least effective teachers would yield very large income gains to affected students.[26]

The large variation in teacher effectiveness, which the past research indicates is unrelated to salaries, indicates that different schools can spend the same amount while getting very different student outcomes, depending on their ability to hire more or less effective teachers. Because teacher salaries and benefits are on average 58 percent of total current expenditure, hiring and retention decisions for teachers can make a substantial difference in student achievement.[27] This conclusion strongly reinforces the conclusion that how money is spent is a key issue, since the quality of the teacher stock in a school or district can differ dramatically when spending does not.

5 Resources and Outcomes (Contemporary Studies)

There is a sharp break between recent and historical research on elements of educational production. The prior analysis most often attempted to estimate the marginal contributions of a variety of purchased inputs and other inputs (families, peers, etc.) as depicted in the production functions of Eq. 1 and 2. These observational studies employed administrative and survey data on school operations and were most focused on understanding the overall impacts of schools and other factors and less focused on identifying the causal impact of specific school inputs. With increasing force in the 21st century, economic research has emphasized the causal identification of various treatments (see Panhans and Singleton (2017)). Building upon ideas of randomized controlled trials (RCTs), the empirical methodology has moved to emphasize natural experiments that might provide evidence about the impact of various specific policies. A key element of these studies is a clear change in emphasis of educational research. The more recent analyses concentrate on much more specific programmatic differences where quasi-experimental methods offer the possibility of clearer identification of causal impacts without attempting to define the overall production possibilities.

We review several different subsets of this research most directly related to the prior observational studies. We make an effort to compile the universe of available evidence in each area including relevant international analyses, although it is quite possible that we have not located all of the international studies. While some judgment is required, we searched for subsets of research pursuing

[26] This conclusion holds even if such replacement policies feedback into higher salaries that recognized increased risk (Rothstein (2015), Chetty, Friedman, and Rockoff (2014)).

[27] See expenditure data in Cornman, Phillips, Howell, and Zhou (2022). The implications of retention policies surrounding teacher layoffs illustrate how policies based on teacher seniority can yield dramatically lower student achievement compared to policies based on effectiveness (Boyd, Lankford, Loeb, and Wyckoff (2011), Goldhaber and Theobald (2013)).

quasi-experimental approaches and meeting modern quality standards. We specifically focused on studies that relate to school finance funding, capital projects, class size reduction, and teacher incentives for student outcomes. The appeal of these quasi-experiments is that under certain conditions they can provide unbiased estimates of the impact of specific programs without having to know and to measure all of the other potential factors impacting student outcomes. A central element of this research is an explicit description and justification of the counterfactual, or what would occur without the specific program under consideration.

Of course, nothing comes for free. First, any single estimate of the treatment effect, even if unbiased, does not have to be very close to the impact parameter of interest. Individual point estimates of the key program parameters will equal the true parameter plus sampling error, and the sampling error can be large. Second, while the impact of a program may be reliably estimated, information about the mechanisms underlying the impact may not be produced, making policy application difficult. Finally, while the impact parameter may be well-estimated in the specific circumstance, it may be difficult to know whether it generalizes to other circumstances and whether the results can be extrapolated.

An overarching aspect of these studies is also that they are best suited for studying discrete policies that have immediate short run impacts. Thus, policies that operate through longer run behavioral changes – such as trying to change the character of students entering teacher training by adjusting certification requirements or overall salaries – cannot be easily addressed by quasi-experimental methods. Similarly, general equilibrium outcomes of policy changes are difficult to assess.

We begin with a discussion of the search procedures used to find the relevant set of studies. We then turn to a description and compilation of impact results across alternative programmatic areas. Most attention is given to overall spending results since this work related most closely to school finance decisions both in legislatures and in the courts. We subsequently expand this analysis to consider alternative policies including facilities and capital investments, class size reduction, and introduction of salary incentives for teachers. These latter areas have received considerable recent attention and potentially provide indications of mechanisms for school improvement. We finally provide a discussion of the generalizations and policy implications of these quasi-experiments.

5.1 Study Selection criteria

The analysis employed a structured search of a wide variety of sources, and then systematically eliminated papers that did not meet a series of criteria for relevance and quality. The first step was to conduct the search for journal articles published between 1999 and February 2022 using two search engines that cover the economics and education literatures, respectively: EconLit and the Education Resources Information Center (ERIC). For each of the four strands of analysis elaborated on here (school spending, capital expenditure, class size, and performance pay), we include the search term “education” along with a strand-specific set of keywords as listed in Electronic Appendix Table A1. This search was also conducted for several repositories of relevant working paper series: National Bureau of Economic Research (NBER); World Bank Policy Research; the Institute for the Study of Labor (IZA); the Center for Economic and Policy Research (CEPR); and the CESIfo Research Network. The search was conducted between December 2021 and February 2022, so papers published after this time are not included. We

reviewed the abstracts of the English language articles and selected those papers whose abstracts met three criteria: 1) discussion of a quantitative causal analysis; 2) relevance to one or more of our four strands of analysis, and 3) mention of effects on student outcomes, including test scores and various measures of attainment such as dropout rates, years of education, graduation rates, etc.

Among the studies whose abstracts met these criteria, we then more closely reviewed their econometric methodology and selected those papers whose estimation strategies included sufficient treatment of possible omitted variable or endogeneity bias. These papers included those employing a randomized controlled trial (RCT), difference-in-differences (DD) regression, fixed effects (FE) estimators, regression discontinuity (RD) design, instrumental variables (IV), and variations on these methods.

We then examined the networks of additional papers identified in our first round of search. This included identifying papers cited in the reference list of first round papers and papers that cited first round papers (as identified using Google Scholar's "cited by" feature). The abstracts and methodologies of these studies were then parsed for relevance and quality just as with the first set of studies. This network identification and subsequent filtering was then repeated with the second set of studies. Finally, we further narrowed the pool of studies by ensuring the inclusion of inputs relevant to producing comparable parameters for each strand.[28] For papers examining the effects of school spending, those that do not provide either the base levels of per-pupil spending or the necessary inputs to calculate these levels are excluded. Studies that only provide effects of various policies on gaps in achievement or attainment (e.g., between white and black students or between low SES and high SES students), as opposed to levels, are likewise excluded.

Within a study that provides multiple estimates, the most general estimates were selected. For example, if an author presents estimates for math test scores, reading test scores, and scores pooled across both subjects, we selected the pooled estimates. If an author presents estimates using multiple specifications, we selected the preferred specification as identified by the author. If the author does not specify the preferred specification, we selected the specification that is most comparable to that of other studies in the literature. The details of parameter selection from individual included articles are found in Electronic Appendix Table A2.[29]

In this review, we separate studies of operating budgets and finance programs from those directed at specific inputs to the production process such as capital expenditures for school construction or renovations and class size reduction. Each of these specific input studies obviously involves school spending, but we divide and analyze separately studies about individual mechanisms behind funding changes.

We are most confident that we have found the universe of relevant studies that exists in the published literature and that provides evidence for U.S. schools. We have found additional unpublished studies in major working paper series, but this portion of the search almost certainly has missed other articles that do not appear in these restricted working paper series. Published studies offer the quality

[28] Studies that could not be compared in terms of the outcomes are excluded.
[29] See https://data-nber-org.stanford.idm.oclc.org/data-appendix/w30769/Appendix_tables.pdf or http://hanushek.stanford.edu/publications/us-school-finance-resources-and-outcomes.

assurances associated with peer review. The same is not true for the unpublished studies. More importantly, working papers that have not been published over long period of time raise quality questions. While we include results from other schooling systems around the world, including from developing countries, we emphasize the U.S. results. The different international contexts lead to increased concern about how to generalize from very different institutional frameworks.

The possibility of publication bias introduces one important caveat for our compilation of existing studies. The outcomes of an analysis in terms of the sign, size, and statistical significance of key parameters have, by past observations and analyses, had some influence on publication. This issue has been analyzed in a wide range of disciplines, and it has been found quite broadly to be a serious issue (see, for example, Nissen, Magidson, Gross, and Bergstrom (2016)). The problem has been linked both to the choices made by researchers and the choices made by journal editors. An early study of clinical trials using RCTs found that negative results systematically led to a lower probability of the findings being written up and submitted. In another study, Franco, Malhotra, and Simonovits (2014) analyze a cohort of NSF-sponsored projects in the social sciences and find that "Strong results are 40 percentage points more likely to be published than are null results and 60 percentage points more likely to be written up."

A particular form of the publication-induced incentives is what has been labelled "p-hacking." Head , Lanfear, Kahn, and Jennions (2015) conclude that "A focus on novel, confirmatory, and statistically significant results leads to substantial bias in the scientific literature. One type of bias, known as "p-hacking," occurs when researchers collect or select data or statistical analyses until nonsignificant results become significant." P-hacking has been the subject of recent analysis (and controversy) in economics, but its existence seems indisputable (Brodeur, Cook, and Heyes (2020, 2022),Kranz and Pütz (2022)).[30] While a variety of tests and corrections for publication bias have been proposed, we focus just on presenting the author(s)'s findings from both published and unpublished studies where possible.[31] We also make no adjustments for any subsequent critiques of included works.[32] Nonetheless, the distribution of results is likely affected by these publication issues but to an unknown degree.

5.2 Impact of Current School Spending

The historical research gave little support to the idea that common school resources were consistently related with student outcomes, but this answer came with the obvious questions about interpretation of the observational results. Specifically, the prior observational analyses provided little support for systematic school improvements related to just providing added funds and thus moving the budget constraint out. But, as noted, the studies addressing the effectiveness of spending per se in the early research were subject to significant biases, making them the lowest quality of the various resource investigations. Debates about this portion of the early research led to a public discussion of these issues

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

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

[32] Some critiques are implicitly included when given studies are motivated by concerns about other included studies. There are also studies that replicate some prior work because of questions about the results, but we do not include separate estimates from these (e.g., the critique of Jackson, Wigger, and Xiong (2021) by Goldstein and McGee (2020)).

that often has been, somewhat misleadingly, characterized as addressing the question "does money matter?" (see, for example, Burtless (1996)).[33]

The number of studies designed to understand the impact of spending per se increased during the 21st century. The general evaluation methodology behind many of these lends itself to understanding the impact of various types of expenditure and resource changes. We report estimates from a broad range of studies that examine the effects of increasing per-pupil funds to K-12 schools through policy changes, grants, revenue limit votes, and several other means.[34]

We located and analyzed 43 estimates of the effect of school spending on student outcomes that meet our inclusion criteria. Table 9 provides an overview of the individual studies investigating the impact of school spending on student outcomes. We describe the available studies by methodology, measure of outcome, and published v. unpublished. We also distinguish between the studies of U.S. schools and those from elsewhere. In total, 36 of the 43 estimates stem from studies examining the effect of school spending in the United States. The majority of the estimates are also from studies published in peer-reviewed journals, and of the 11 estimates from unpublished studies, 10 cover spending in the United States. There are 23 estimates of the effect of spending on test scores, and just 2 studies that look at proficiency or pass rates. The 18 studies on educational attainment include 8 estimated effects on high school graduation rates, 6 on college attendance, and 4 on dropout rates. Though the exact estimation methods vary widely between these studies, the most common broad methods include instrumental variables (IV), regression discontinuity (RD), and difference-in-differences (DD). The 4 randomized controlled trials were all performed outside of the US, and other papers applied difference-in-differences or fixed effects techniques. From this overview, it is clear that the studies cover a wide range of circumstances and pursue a variety of identification strategies. We briefly highlight a few of the studies that show the range of approaches taken to estimating the impact of funding differences.

We begin with a set of studies exploiting funding changes related to court actions and then turn to a broader set of studies of funding changes. Importantly, while the results of various resource studies have entered in school finance court deliberations, these court cases themselves have been an important catalyst and source of variation for the study of public K-12 funding. We discuss a subset of these studies followed by a sample of other funding change experiences that have motivated resource impact studies.

The various underlying financial reforms, both judicially-based and nonjudicially-based, potentially introduce sharp departures from the prior spending distributions and suggest a potential source of exogenous school funding that can support well-identified analysis of the impact of variations in funding. The key element in all of these studies remains, however, the necessity that the spending increases are not correlated with other separate policies or actions that impact student outcomes, otherwise we would have biased estimates of the impact of spending. In order to influence student

[33] Many researchers correctly said that differential funding of schools may or may not have shown impacts within their samples, but few argued than zero money or even that cutting back on funds would have no impact. At the same time, the "money does not matter" language has appealed to both the media and various policy advocates seeking specific policy solutions.

[34] Note that our review of recent spending studies is not the first. Jackson and Mackevicius (2021) conduct a review of both spending and capital studies, although that study differs significantly in approaches to the analysis.

outcomes, spending must involve some set of policies. If these are just the policies that are employed to implement a new spending program, they will not bias the reduced form estimates of the impact of the spending. On the other hand, if there are concurrent policies that are correlated with the spending being analyzed but are not the direct result of the new spending program, the impact estimates will be biased estimates of the reduced form spending parameters.

Jackson, Johnson, and Persico (2016) investigate the impact of school spending on longer term outcomes using data from cohorts born in 1955-1985. They introduce the idea that court ordered spending could provide exogenous variation that permits causal estimates of the effects of additional educational spending. They consider court decisions in equity cases in 28 states in the 1970s and 1980s and follow the long-term effects of funding changes on cohorts through 2011. To isolate the effect of spending from confounding factors, they construct the predicted reform-induced change in spending for each exposed district based on other districts in the sample. Using this measure for variation in spending, they estimate the effect of exposure to increased education spending sustained over 12 years on student outcomes including high school graduation rates, adult poverty levels, and adult wages. They leverage the variation in the timing of passage of the reforms in a difference-in-differences framework with instrumental variables to compare the difference in outcomes between affected and unaffected cohorts. Their estimates imply very large impacts of increased spending, particularly for low income students. For example, they estimate that "a 22.7 percent increase in per-pupil spending throughout all 12 school-age years for low-income children is large enough to eliminate the education gap between children from low-income and non-poor families" (Jackson, Johnson, and Persico (2016), p. 26).

Candelaria and Shores (2019) study the more recent reforms taking place in the "adequacy era" of court actions. In these, proponents argued that the states have an obligation to provide some minimum level of funding appropriate for providing an adequate education. This study considers school finance reforms implemented between 1989 and 2010 because of court orders, applying a difference-in-differences framework that aims to account for heterogeneity across district poverty levels. They estimate that the highest poverty quartile, which experienced an 11.5 percent to 12.1 percent increase in per-pupil spending seven years after reform, had a 6.8 to 11.5 percentage point increase in graduation rates.

A third example of a study using court actions as a source of exogenous variation in spending is Buerger, Lee, and Singleton (2021). They examine the effects of funding reforms on reading and math NAEP scores for students in 4th and 8th grade, focusing their analysis on the role of state accountability measures as potential mechanisms for improving the efficiency of spending. They implement an event study DD approach to find that reform-induced increases in educational spending[35] averaging 7-9 percent of base spending in low-income districts led to 0.012 SD increases in test scores when paired with accountability measures in place and .006 SD test score gains in districts without accountability measures.

[35] They rely on the identification of reform that comes from both court and legislative actions as defined by Lafortune, Rothstein, and Schanzenbach (2018).

These three studies are interesting because they employ common data and methodologies, yet it is difficult to judge whether their results are consistent with each other or not. Each concludes that the spending generated by a subset of school finance court cases appears to yield positive student results. But that results by itself is minimal justification for most policy options. We return to this in the next section after considering a second set of studies with a common basis.

While this set of studies focuses on court-ordered school finance reforms, another strand of the literature focuses on contexts in which the variation in school spending was driven by legislative action. One such example is Michigan's Proposal A, a reformulation of Michigan's school funding approach that prioritized a shift away from property taxes and established a foundation level of funding designed to provide an adequate education for all students. Michigan's Proposal A, which was approved by a public vote, led to a rapid growth of budgets in low-spending areas. This policy is examined by Papke (2008), Roy (2011), and Baron, Hyman, and Vasquez (2022).

Papke (2008) uses district-level panel data from 1992-2004 to investigate how funding affects student pass rates on Michigan's statewide 4th grade math exams. Allowing for a rich lag structure in school spending, district fixed effects, and IV estimation, she estimates a 3.7 percentage point increase in pass rates for each additional 10 percent increase in average real spending with the effects being much larger in districts with below average initial pass rates. Roy (2011) conducts a similar analysis that also uses the funding reform as an instrument for school spending in the estimation of impacts on 4th grade math and reading test pass rates in 1998-2001. He estimates that a $1,000 increase (on a mean base spending of approximately $5,000) would increase reading pass rates by 3-6 percentage points and math pass rates by 6-8 percentage points.

Baron, Hyman, and Vasquez (2022) also study Michigan Proposal A along with a set of bond referenda for capital expenditures to compare the effects of additional operational and capital spending. They use a two-stage least squares framework to examine the effects of extra funding induced by Proposal A and sustained over grades K-3. Their sample consists of students in kindergarten in 1995-2004, and they find that a 10 percent increase in funding over these first four years of schooling leads to a 12 percent of a SD increase in test scores in the short term, a 3.4 percent increase in high school graduation rates, and a 4.3 percent increase in college attendance. Broadly, their analysis reveals that exposure to higher operational and capital spending in grades K-3 lowers the likelihood of adult arrest through higher educational attainment and improved noncognitive skills. Test scores are not impacted in the long term.

While the first two Michigan studies of exam pass rates appear to provide similar estimates of impacts, it is not clear how the results of the third study relate to them. The last study finds a variety of impacts in different areas but also finds no long term impacts on student achievement.

The common thread of these studies is employing a distinct change in spending to identify the causal impact of spending on measured student outcomes. When looked at in more detail, there are three primary conclusions from this sample of studies (and the larger set of related studies). First, both the outcomes of the analysis and the measurement of changes in the budget constraints vary widely across the studies. Second, the studies employ very different analytical approaches to the analysis. Third, and perhaps most important, the effects that are estimated in each study come within various

restrictions and programmatic requirements that may or may not strongly influence the resulting impacts.

Creating comparable parameters

Because the analyses and empirical approaches define and measure the fundamental inputs and outputs of the educational process in different ways, we seek to harmonize the measurement so that the estimated impact parameters are as comparable as possible. Because of differences in measurement and reporting of results, this harmonization is not trivial, but it is crucial to obtaining reliable comparisons of different estimates of the impact of spending. Following that, we return to the other issues of comparison.

For each study, we compute the effect of a 10 percent increase in real (inflation-adjusted) per-pupil school spending on standardized outcomes for the general population of students. Because studies do not all report estimates in this form, we scale and transform the estimates provided accordingly, so that we may facilitate more informative comparisons and draw conclusions across various contexts. The general steps to do this are explained in this section, and the steps for standardization of each individual study are provided in greater detail in Electronic Appendix Table A2.

To capture the effects of sustained spending increases on student outcomes, we select estimates taken four years after a policy change or from the beginning of the study period. If this is not available, we take the longest period up to four years. For event study specifications providing coefficients on "years post" a given reform, we use a four-year period as well.

For each study, we collect the average change in yearly spending from the initial levels over the period of study, usually four years as detailed above. For studies that either present policy effects on spending and student outcomes separately or leverage instrumental variables estimates with spending changes as the outcome in the first stage instead of presenting effect sizes in terms of spending changes, we connect spending and outcomes accordingly. Because we aim to obtain externally relevant estimates, we represent the changes in spending as a fraction of the baseline level of per-pupil spending in the sample. [36] We use comparisons of impacts for a 10 percent change in spending. Because of the sharp rise in spending per pupil seen previously, it would be inappropriate to compare simple inflation-adjusted spending levels because the actual date of application, which varies widely across studies, would then be important.

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

It is generally not possible to put studies of pass rates on a scale that is comparable to the estimated impact parameters based on standard deviations of test scores. Proficiency rates depend on

[36] When provided, we use the baseline spending from the year prior to the policy change or the last year prior to the study period. When this level of detail is not provided, we take the average per pupil spending over the study period. Using this ratio, we then scale the provided estimate to represent the effect of a 10 percent increase in base spending levels.

the cut scores chosen by a standard-setting process. Changes in cut scores placed at different points in the achievement distribution can vastly and unpredictably affect the interpretation of impacts (Holland (2002), Ho (2008)). Thus, it is difficult to generalize from any studies using pass rates. When the outcome is a fraction of students above a proficient score threshold (i.e., a pass rate), we still report results on the percentage change in passing,[37] but we do not attempt to compare the magnitudes of changes to other test score estimates.[38]

In studies where effects are reported separately for different test score subjects, grade levels, or demographic populations, we use the reported standard deviation for that given subgroup if available. If the student-level standard deviation is only available for the full sample, we use this general metric. To convert estimates into student-level standardized units if not already presented as such, we divide the raw effect by the standard deviation.

Some of the original studies focus on school attainment, school completion rates, or the like, which are obviously measures of time inputs into the educational process. They are also frequently used as outcomes when there are no measures of achievement or learning, but they remain crude surrogates for student performance. The recent pandemic underscores the problems with these attainment measures, because school closures plus altered learning patterns make a year of schooling during the pandemic very different from a year of schooling outside of the pandemic period.[39] But this is a more general problem because the quality of schooling varies over time and across space. We translate the attainment measures into percentage change measures. For dropout rates, we multiply the original effect by -1 to make the impact more comparable to other measures of attainment, in which positive estimates imply desired impacts. At the same time, much of the research and policy discussions generally treat concerns about high school dropouts as qualitatively different from college attendance – making aggregation of these impact parameters problematic.[40]

As noted earlier, when authors provide estimates across various test subjects and pooled estimates, we will take the most general specification. If authors do not provide pooled estimates, we

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

[38] The previously discussed analyses of Proposition A underscore the problems. While Papke (2008) and Baron, Hyman, and Vasquez (2022) provide internal confirmation of the positive impact of the Michigan finance changes on low income districts, it is not possible to place the magnitude of any results in the distribution of other estimates of impacts on achievement.

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

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

average across test score subjects by computing the simple average of the effects.[41] Similarly, we report the estimates from the most general specification with regards to sample composition. If authors only provide separate estimates across grade levels, income levels, race, etc., we compute average estimates using a precision-weighted mean to combine estimates across grade levels. To combine estimates across populations with different demographic characteristics, we weight estimates with the relative share of their respective subgroups in the overall population.[42] Applying these steps to each study as detailed in Electronic Appendix Table A2, we construct a parameter that presents the estimated effect of a 10 percent increase in school spending on student-level standardized outcomes. The set of studies providing estimates of the impact of added spending on student outcomes are shown in Table 10. For each of the identified studies we provide our standardized outcome measure along with the estimation approach and a short description of the source of the estimates.

The importance of standardizing the parameter estimates is directly seen from the studies that address the same policies in Michigan. Papke (2008) estimates that a 10 percent increase in school spending due to the implementation of Michigan's Proposal A led to a 5.9 percent increase in test pass rates, while Roy (2011) finds that pass rates increased by 5.4 percent. These estimates line up quite well, which is not immediately apparent when comparing the results as initially presented by the study authors. Unfortunately, given their focus on pass rates, it is not possible to place these studies within the distribution of impact parameters for achievement.[43]

School spending and achievement

Our main focus is U.S. studies that measure the impact of spending in terms of student achievement. We provide information on the international studies for comparative purposes, but the varied educational systems preclude obvious ways to generalize to U.S. schools. The other outcome measures beyond achievement provide a broader view of outcomes, but they are also less reliable measures of the learning and skills from schools. While we include the estimated impacts from the unpublished studies, they have yet to be fully vetted by the journal refereeing process. Thus, we lean more toward the published papers that have already received a thorough peer review.

Table 11 summarizes the 43 studies by outcome measures investigated. It also separates the 36 estimates for U.S. schools in the bottom panel, the focal point of our analysis.[44] Of the 16 U.S.

[41] We compute the standard deviation of these average effects by following Chapter 24 of Borenstein, Hedges, Higgins, and Rothstein (2021) and assuming a correlation of 0.5 among test score subjects within the same grade as done by Jackson and Mackevicius (2021).

[42] For estimates across various grades, we follow Chapter 23 of Borenstein, Hedges, Higgins, and Rothstein (2021) to use an assumed correlation of zero as done by Jackson and Mackevicius (2021). To combine estimates across demographic subpopulations, we apply the methods outlined in Borenstein, Hedges, Higgins, and Rothstein (2021) in Chapter 24.

[43] Note, however, that the third study of Michigan impacts is readily included. Baron et al. (2022) study the effect of spending on both achievement and attainment in this same context, finding that a 10% increase in funding yielded test score gains of .011 standard deviations, a 3.4% increase in high school graduation rates, and a 4.3% increase in college-going.

[44] Note that the seven non-U.S. studies include three from developed countries and four from developing countries.

achievement outcomes, 14 estimates are positive, and 9 of these are statistically significant at traditional levels. The overall median effect size for a 10 percent spending increase is 0.07 standard deviations.[45] With a few exceptions, the estimates near the median are the most precisely estimated, but the range of estimates is startling. The estimates of the test score change in standard deviations from the increase in spending of 10 percent go from -0.244 (and not statistically significant) to 0.543 (and statistically significant). Figure 7 summarizes the distribution of the estimated effect sizes of a 10 percent increase in school spending on student achievement outcomes where the included studies are restricted to U.S. schools. We plot the standardized effect size along with the 95 percent confidence interval in a forest plot for test scores (in SD) only.

The estimated impacts of federal compensatory education funding provide an example of the contrasting pictures of the impact of resources and provide an introduction to some of the interpretative issues. Title I, which was the largest component of the 1965 Elementary and Secondary Education Act, provided a large infusion of federal funding into public education. Its contemporary successors, the 2001 No Child Left Behind Act and the 2015 Every Student Succeeds Act continue to provide the largest source of federal funding for K-12 education. The aim of the program was the provision of funding to low-income and otherwise disadvantaged students. Cascio, Gordon, and Reber (2013) found that exposure to 10 percent additional funding as a result of Title I grants led to an 18 percent reduction in 18–19-year-old dropout rates between 1960 and 1970. Johnson (2015) finds that a 10 percent increase in county-level funding led to a 12.9 percent increase in graduation rates, which was concentrated in poor children. These findings on graduation rates are of the same order of magnitude as those of Cascio, Gordon, and Reber (2013) on dropouts. The estimated percent effects are higher on dropout rates, which may partially reflect the differences in sensitivity of the two measures. On the other hand, Weinstein, Stiefel, Schwartz, and Chalico (2009) investigate the impact of federal Title 1 spending for poor children on the performance of students in New York City using a regression discontinuity design with panel data and fixed effects. They find that effects for all measures of academic performance are negligible, with a 10 percent increase in funding leading to *decreases* in test scores by .08 standard deviations for both math and reading, though this finding is not statistically significant.

The small number of replications of estimates for similar or identical policies precludes strong conclusions.[46] Nonetheless, they help to sort among the possible explanations of the variation in impact estimates. Similarity in replication, where possible, implies that the effect of pure sampling error may be small relative to differences in policy environment (ρ) or in study quality across the full range of studies. We return to the interpretive issues below.

Table 12 divides the results by methodological approach. The test score estimates for the U.S. studies are almost evenly divided across instrumental variable estimates, difference-in-differences estimates, and regression discontinuity estimates, but the resulting array of estimates is very different

[45] If we do a common random-effects meta-analysis on the estimated impact parameters, we get an estimate of the mean impact of 0.066. But, as explained below, such an estimate comes from aggregating across very heterogeneous underlying parameters and has little meaning.

[46] This discussion uses a looser definition of replication than advocated in Clemens (2017). Here we also consider robustness analyses that attempt understand differences in parameter estimates across related circumstances as the original.

across these approaches.[47] The median for RD estimates is very much larger than that for the other two, and the range of estimates in the DD studies is much broader. As the same time, these differences should not necessarily be attributed to the approach per se because the different approaches are used for very different samples and for alternative sources of exogenous variation.

Table 13 illustrates the range of study characteristics that may influence estimated magnitudes of test score effects in the United States. Studies differ in whether the associated source of variation is motivated by a school finance court case. Differing data availability across states and time periods contributes to the heterogeneity in levels of measurement for outcomes and for spending, though almost all of studies measure both at the district level. Finally, just over half of the studies focused on comparing outcomes within a single state, while the rest include several states with their inherently diverse policy environments. Because of the small overall sample of studies, however, we cannot determine whether any differences in median estimates along these dimensions are of economic interest.

To benchmark the collection of estimates, we can refer back to the historical measures of historical unconditional spending impacts on achievement in Table 2. The overall median estimate of the impact of 10 percent added spending from the 16 spending studies in Table 11 was 0.07. Looking at unconditional pre-pandemic NAEP scores in Table 2, we see that the increase in math scores for students in primary and middle school normalized for aggregate spending increases exceed this median estimate.[48] The much smaller unconditional gains (per 10 percent spending increase) for reading are all less than half of the median estimated impact of spending.

The unconditional achievement increases reflect the combined impact of schools and other factors, most notably family inputs. While there are not clear estimates of the causal impact of family factors on achievement, changes over time indicate both positive and negative trends in various measured components, and there is not any strong evidence of large changes in family impacts in either direction.[49] Comparisons of the estimated impact parameters to the unconditional observed test score gains do suggest that the impact estimates above the median are unlikely to reflect estimates that can be generalized to common increases in budgets as opposed to increases in specific spending or particular policy environments. Finally, it is unclear how to interpret the much smaller gains in NAEP scores recorded after the COVID pandemic. While those declines in scores do not reflect normal operations of schools, it is also difficult to ignore the long term patterns of gains that include that period.

As a general rule, the studies of spending changes seldom identify uses of these funds so these aggregate effects are interpreted as observations about the varying impacts of added funds without providing direct guidance about how any funds should be used. Importantly, the movement of budget

[47] Note that Brodeur, Cook, and Heyes (2020) conclude that "our results suggest that the IV and, to a lesser extent, DID research bodies have substantially more *p*-hacking and/or selective publication than those based on RCT and RDD." Subsequent analysis introduced some doubt about these DID results but not the IV results. This concern about the IV estimates is consistent with the critique of Jackson, Wigger, and Xiong (2021) by Goldstein and McGee (2020). It shows that slightly different construction of the instruments used in the IV work produces radically different results.

[48] Math for 17-year-olds, however, shows much smaller increases in effect size for a 10 percent spending increase.

[49] See, for example, the discussion in Hanushek et al. (2022a).

constraints involve a variety of underlying spending initiatives that themselves have been analyzed. We return to differences in these specific inputs below.

School spending and attainment

For the U.S. studies in our sample, 18 focus on attainment, or the quantity of schooling, looking variously at high school completion, school dropouts, and college enrollment. These are clearly crude measures of the learning and skills of students, but they have a long history of use in labor economics. We provide the results of these studies with impacts translated into percentage change in the specific outcome per 10 percent increase in spending. As is evident in Table 11 and Figure 8, all 18 estimated effects for attainment, as measured by graduation, dropout, and college-going rates, are positive, with 14 of these reaching conventional levels of statistical significance. The median impact implies that a 10 percent increase in school spending will increase high school graduation, college enrollment, or another metric of attainment by 5.7 percent.

This median approximation of impact clearly needs to be interpreted with caution, as it is reasonable that dropout rates and college-going rates would not have the same degree of sensitivity to school spending and would be affected through different mechanisms.[50] As with test scores, most estimates lie quite close to this median, but the outliers are dramatically different. At the low end, a 10 percent spending increase yields a 1.8 percent improvement in attainment, while at the high end there is an unbelievable 85 percent improvement in dropout rates.[51] Because of the interpretative difficulties with the attainment measures that arise not only from the pandemic but also from previously identified differences in quality across schools, states, and time, we place much less emphasis on these findings about various components of attainment.

Generalizing from Spending Estimates

The most immediate findings from the previous tabulations are that a large proportion of the estimates of spending impacts are statistically insignificant by conventional standards and that the point estimates are widely different across studies – from negative to very large. These findings are qualitatively similar to the previous production function estimates. Here, however, we need to take a different perspective in evaluating them. If we assume that each of the underlying impact estimates is unbiased, it is difficult to conclude that the range of estimates across studies simply reflects sampling error. If not sampling error, we are left with an explanation either that some of the studies are flawed such that the estimates are biased or that the estimates do not reflect a common impact parameter.

The variation in the estimates can be better understood by first decomposing the set of spending-achievement results into sampling variation and underlying parameter heterogeneity.[52] Using

[50] The typical analysis of college attendance decisions focuses on tuition costs and financial aid (Dynarski, Page, and Scott-Clayton (2023)) or supply considerations (Bound, Lovenheim, and Turner (2010)), as opposed to spending on their secondary schools.

[51] The sample in this latter study (Lee and Polachek (2018)) features a low base dropout rate of 3.056%. Thus, the estimated effect of a .2599 percentage point reduction for a 1% increase in base spending translates to a 2.599 percentage point increase for a 10% increase in spending, which represents about 85% of the base rate. Though it is not immediately clear whether the authors mean for the estimates to represent a percentage change or a percentage *point* change, we infer that they represent percentage *point* changes from the discussion on page 139.

[52] We thank Larry Hedges for suggesting this step.

the standard I^2 measure with our 16 separate parameter estimates (Borenstein, Hedges, Higgins, and Rothstein (2021)), we find that 50.6 percent of the variance comes from between-study variation. While the range of estimated impact parameter is reduced by eliminating the outliers from the Kansas study by Rauscher (2020), their large sampling errors mean that they receive very little weight in the calculation of between-study heterogeneity, and dropping them does not significantly change the estimate of I^2. We are left with half of the estimated variation reflecting variations in the underlying true impact parameters.

In order to interpret the heterogeneity in the estimated impact parameters, it is important to understand the context for each of these studies. They are designed to deal directly with internal validity, which might potentially be compromised by omitted variable bias or by selection and reverse causality issues. As such, under specific conditions the various evaluation approaches provide for unbiased estimates of policy impacts. But, they do so within the constraints of the analysis because they are conditional upon the institutional and sampling background of the study. In other words, the impact parameters estimated for the funding-achievement relationship are dependent upon the setting in which they were estimated. The generalizability of any findings will depend upon the potential impact of these factors when translated to a different environment.

The key to generalization and interpretation of the results is ensuring that the parameters being aggregated come from similar institutional circumstance and populations both with respect to others in the analytic grouping and to the circumstance to which the aggregation will be applied. In other words, the interpretation of aggregation of the results across studies will depend on the institutional and environmental influences that are held constant in the underlying analysis. As an analogy, combining the results from a series of randomized control trials across closely related age groups about a specific type of COVID vaccine may provide informative results that generalize to even larger age ranges. But aggregating impact results from RCTs for mRNA COVID vaccines with those from traditionally-produced vaccines for seasonal flu may be less useful. Just reporting the average impact of a flu-related vaccine would not be very relevant for a policy to deal with the spread of COVID-19 or with the common flu.

As evidenced by the large portion of variation in estimated impacts stemming from between-study heterogeneity, the diversity of contexts in the underlying studies of school spending make it difficult to draw externally valid conclusions about funding changes or to compare estimates across studies. First, the settings differ wildly. Several included studies measure the effect of Title-I related schools spending (Cascio et al., 2013; Johnson, 2015; Weinstein et al., 2009), which necessarily limits the sample to schools with lower income students and restricts the potential programmatic elements of the subsequent spending to follow Title I regulations. Others compare impacts across states, presuming that the policy differences across states do not matter – contrary to the prior results about the biases of aggregation described in Table 7. Indeed, Buerger, Lee, and Singleton (2021) explicitly show that state differences in accountability policies – one of many policy differences – interact significantly with the effectiveness of added resources. The heterogeneous conditions that prompted each policy change also yields differential estimated effects. Spending equalization reforms studied nationally (Brunner, Hyman, and Ju (2020), Buerger, Lee, and Singleton (2021), Candelaria and Shores (2019)) and in individual states (Guryan (2001), Hyman (2017), Papke (2008), Roy (2011)) aim to increase the resources flowing to

districts that spend below what is deemed a minimum basic level for an adequate education. It is reasonable to expect the effect of these policies to differ from that of a policy in which extra per-pupil windfalls arise due to peculiarities in state funding formulas that are not necessarily targeted upon any observable district-level characteristics (Miller (2018)). The included studies also cover a broad range of time, covering data from 1950-2018. We do not find that the measured impacts of spending on student outcomes vary systematically with time, even though prevailing views of optimal school policies may differ over time.

The results also identify the outcome impacts of spending in a wide variety of situations that differ significantly from anything resembling a simple move outward of the budget constraint. For example, during a time of budget reduction, a school district may have to layoff teaching personnel, but who is laid off is generally prescribed by teacher contracts and state laws (e.g., Boyd, Lankford, Loeb, and Wyckoff (2011), Goldhaber and Theobald (2013)). Therefore, it is difficult to think of the estimated impact of budget declines (e.g., Jackson, Wigger, and Xiong (2021)) as capturing the same spending parameter as that from added resources from the diverse set of court decisions (e.g., Jackson, Johnson, and Persico (2016)).

The estimates that rely on spending variation associated with court cases are motivated by the idea that the spending after court decisions can reasonably be assumed to be exogenous to other factors that might influence outcomes. But the magnitude of any spending response or even whether or not a state responds to a court decision is likely to vary with other political and institutional factors in a state, complicating the identification of the spending impact parameter and the generalizations that can be made.[53] These problems are magnified by combining court-imposed spending directives and legislative actions not resulting from a court decision (Lafortune, Rothstein, and Schanzenbach (2018)). It is difficult to think of the legislative actions as being random.

Even more importantly, the interpretation of spending parameters estimated from samples that cross states is difficult. These average impacts across multiple states incorporate the impact of very different regulatory and incentives arrangements. The significant impact of state policies on bias of funding impacts shown in Table 7 suggests that the estimates merging the impact of multiple state educational structures are difficult to interpret. As seen in the historical studies, the size and impact of funding changes is likely to be affected by a series of institutional features of the schools of each state.

In sum, it is difficult in general to interpret the set of estimates displayed in Table 9 as relating to a common spending parameter. Each of the estimates of the impact of spending is conditional on a series of underlying restrictions and institutional characteristics that guide the usage of any additional funds. We provide information about the median estimate, but the variations in estimates are central to any interpretation and policy use. Indeed, the variations in estimated parameters may provide insights into the mechanisms that could lead to larger impacts from spending variations by ensuring that funds

[53] As a general rule, courts avoid specifying the amount of spending changes that a decision would require, because only the legislatures in the states have the power to appropriate funds. Moreover, historically some legislatures have not responded to court rulings that directed increased spending. These choices are obviously not random.

were used in the most productive way. Unfortunately, given the relatively small number of separate studies and the current lack of good descriptions of the relevant educational institutions and policies, this remains an open question.

Because the emphasis in individual studies has been more on the identification of a specific spending parameter, less attention has been given to comparability to other estimates of impact parameters. The wide differences in estimates of the spending impact parameters could come from a variety of sources. We have set aside any issues of study quality including publication bias and p-hacking concerns, although these issues undoubtedly enter. The central issue in interpretation is that each of these studies – while focusing on internal validity – is conducted with a specific sample and a particular set of policies that produce arguably exogenous variation in spending, and the impacts reflect the responses of decision makers both to the spending changes themselves and to the institutional and environmental structure. We refer to this combination of factors simply as "how resources are used." "How" thus wraps together the impact of the institutional and regularity structure and the policies and decision maker actions that come to play in implementation of any increased resources.

The previous results imply that the key to effective policy is understanding how resources are effectively used and what leads to ineffective uses. The varied strength and significance of the estimated impact parameters indicate that how resources are used is crucial to the success of any potential funding program.

It is not possible to get around these issues by aggregating across the existing evidence from very different underlying institutions and policies. Many of the estimates reflect very specific policy changes that involved constrained spending options and that are very different from changes in the overall budget constraint. Less obviously, apparently unconstrained spending changes such as coming from various state school finance cases are subject to varying constraints. Prior cross-state analyses highlight the importance of state policy differences in the estimation of spending impact, leading to potential estimation problems even in the context of modern empirical approaches. In simplest terms, meta-analytic approaches offer ways to improve on individual estimates when the studies being aggregated fall into a reasonably similar class of institutional and policy environments. These approaches then provide improved policy information when applied to the same similar institutional environments. Unfortunately, the existing research has provided little clarity on appropriate classes of institutional and policy environments, and the limited number of existing studies constrains the generalization of any specific results.

5.3 Spending and the Impact of Specific Inputs

The prior discussion looked at evidence related to changes in the budget constraint of schools and student outcomes measured in different ways. A major conclusion was that the studies gave very different estimates of whether overall resources consistently lead to improved outcomes and of the magnitude of any changes.

Even if the answers were more consistent, a policy dilemma remains. There is no real description from these studies of what mechanisms are most likely to lead to significant improvements

in student outcomes. We therefore turn to related work that focuses on the role of specific inputs. In this, we consider capital spending, class size, and teacher incentive programs. Each of these input-related investigations is amenable to well-identified empirical analysis, and each follows policy changes that have had considerable traction in the U.S.

Capital Spending

The exploration of the effects of school capital spending such as that of building renovations or new school construction is a relatively recent concentration in the literature, reflecting in part their amenability to causal identification. While capital expenditures and interest on debt have varied with demographic changes and population growth over time as a proportion of total expenditures, they have been stable at slightly over 10 percent in recent years. Finance and budgeting for these expenditures differs significantly from current operating expenditures, and these expenditures almost always follow different procedures and decision making processes. Because large capital expenditures are generally funded by long term borrowing and long term commitments for taxes to fund them, states typically require initial voter approval of projects. Permissible uses of any resultant capital funds are obviously highly constrained.

Summarizing these studies of capital expenditure poses a unique challenge to researchers and policymakers. Local capital expenditure projects have varying purposes ranging from repair and replacement of dilapidated buildings to meeting demands of local population growth to nonacademic purposes such as enhanced sport facilities and to providing new equipment such as school buses and new computers. Moreover, while often involving lumpy expenditures, these projects have varying construction periods and different useful lives. As such, it is difficult to compare directly the exact nature of the expense and relevance of differing capital projects, particularly as found in evaluations of different specific programs.

We find 20 estimates of the effect of capital expenditure in student outcomes and detail these in Table 14. All but one of these estimates considers impact effects in the United States. Twelve of these estimates come from regression discontinuity designs leveraging close elections for school district bond referenda supporting capital expenditures. Of the remaining eight estimates, three come from instrumental variable designs, and five are derived from various forms of difference-in-differences or fixed effects specifications. The first of these studies was published in 2010, and the underlying sample data span 1987-2014.

To get some flavor of these various programs, we describe two sets of closely related studies. These not only demonstrate the character of different approaches but also show the sensitivity of the findings to specific analytical decisions. The studies utilizing RD designs with close elections for school district bond referenda examine capital expenditures in California, Michigan, Wisconsin, and Texas. Of particular interest are studies by Martorell, Stange, and McFarlin (2016) and Schlaffer and Burge (2020), as both investigate bond elections in Texas. In this context, funding school construction and renovation projects through the issuance of bonds requires securing a majority vote from local taxpayers. It is assumed that districts whose electorates narrowly approve the proposed capital expenditures are very similar along unobservable dimensions to those that narrowly reject the proposals and thus do not proceed with renovations and construction. The plausibly exogenous placement above or below the vote cutoff for passage is used to estimate the effect of each referendum.

Martorell, Stange, and McFarlin (2016) employ data spanning 1997-2010 to investigate the impact of capital spending projects in Texas. The average project costs $10,300 per pupil (in $2022), which is typically spread over several years and represents 600 percent of the average yearly district-level per-pupil capital spending. They examine the effects of bond passage on attendance and standardized test scores for students in grades 3-8 using two strategies, an RD design using bond passage as treatment and an event study using imputed renovation and opening completion as treatment. An RD analysis of the impacts of bond passage on capital investments suggests that capital expenditure doubled in the first two years for districts that approved their projects, that the share of students attending new school buildings doubled, and that the average age of school buildings was reduced. Looking to student outcomes, they find positive but statistically and economically insignificant effects of bond passage on test scores using the now-standard dynamic RD design to account for multiple elections proposed by Cellini, Ferreira, and Rothstein (2008). The gap between baseline low- and high-scoring students is unaffected by the facility upgrades and new construction, though some test score gains emerge by year six after bond passage for poorer students.

Schlaffer and Burge (2020) find contrasting stronger effects. They also perform an RD analysis of the impacts of school facility upgrades and construction on students in Texas also using close bond elections as a source of variation. They focused only on votes that generated new facility construction and only on students that do not change schools after bond passage. Examining elections carried out between 1997 and 2014, their estimates suggest that narrow bond passage has a positive and significant effect on both math and reading test scores for students in grades 3-8, and they find that gains are larger for students in the lower end of the achievement distribution. Using a modified RD approach, they find that bond passage raises scores by 0.06 SDs after six years. Looking directly at the openings of new schools, they find gains of 0.1 SDs.

Both studies use detailed Texas data applied to the same setting along with a commonly accepted approach to investigate the causal impact of capital expenditure on student outcomes. Both have sufficient power to detect plausible impacts. Yet, the results are strikingly different, suggesting that more than just sampling variation is at play.

Of the studies using alternative estimation methods, two cover an Ohio capital subsidy project. Created in response to a 1997 State Supreme Court case ruling regarding the need for equitable sources of funding for school construction projects, the Ohio School Facilities Commission provides state-sponsored subsidies for school facility upgrades. Districts are ranked based upon property values and income, and a ranking cutoff for eligibility is established each year. The cutoff moved up every year, with more wealthy districts becoming eligible with each new year. The amount of local funding that a district must provide for its project is commensurate with its ranking; poorer districts are more heavily subsidized. This project disbursed over $10 billion between 1997 and 2011 for upgrades in 231 districts.

Conlin and Thompson (2017) utilize this Ohio program to examine the relationship between capital spending and student outcomes. They measure the effects of capital spending and lagged capital stock increases on student performance on state math and reading test proficiency using an instrumental variables approach. This method exploits a first stage relationship between capital expenditures and stock and school district eligibility for the subsidy in the current and prior years (as well as in the three years prior) as established by a wealth ranking and yearly cutoff. Given that this cutoff gets more lenient each year, the timing of eligibility varies by district. In their preferred

specification, they also use a first-difference estimation for the first stage, addressing unobservable time-invariant district-level characteristics. With district-level data on spending, test scores, and demographic composition for 1997-2011, Conlin and Thompson find that capital expenditures can harm student performance at first, with negative effects in the year of spending and in the year after. They also find that increases in capital stock can lead to improvements in student performance 3 or 4 years later, which they cite as evidence for the ability of completed capital projects to aid in student outcomes.

But, Goncalves (2015) earlier studied the impact of the same Ohio School Facilities Commission's program of funding school construction on math and reading scores and home prices. He uses test score data from 2005-2014 to examine the lagged impacts of exposure to both construction and completed capital projects. This is done through a fixed effects approach that includes district and time fixed effects and eligibility group-specific year effects, which restricts the comparisons to within groups of districts that became eligible at the same time. Each students' treatment is determined by the time they spend exposed to construction activities or a completed construction project. Using this model, Goncalves finds that all students are negatively impacted during construction, as evidenced by deleterious impacts on scores in both math and reading. Unlike Conlin and Thompson (2017), he finds no statistically significant positive effects of capital projects post-completion. Goncalves notes that these effects are not uniform. That is, he finds that the negative effects are concentrated in the poorest quartile of districts and in middle/high school age students.

These pairs of studies are each arguably replications of analyses of the same input parameter, but the quite disparate results for analyses of impacts in the two identical treatment situations for Texas and Ohio raise significant questions of interpretation.[54] These differences point either to large sampling errors or to more fundamental problems with the identification of the spending parameters.

Because of the inherent underlying heterogeneity of the broader set of capital project evaluations listed in Table 14, we present the estimates of the effects of capital expenditures as they are reported in their respective studies, only scaling by student-level standard deviations and combining across test score subjects and grades when necessary. We do not attempt to construct identical spending parameters, but the findings in Table 14 provide an image of the distribution of findings in the literature.[55] To capture the effects of completed construction on student outcomes, we select estimates taken six years after a bond referendum or from the beginning of the study period. If this is not available, we take the longest period up to six years. Exact timing is detailed for each estimate in Electronic Appendix Table A2. We scale each test score estimate by the student-level standard deviation and each pass rate or attainment estimate by the mean baseline rate as done with the standardized school spending parameters earlier in this section.

[54] As a general rule, however, a single replication is unlikely to provide strong evidence about an estimated impact parameter (Hedges and Schauer (2019)).

[55] Jackson and Mackevicius (2021) attempt to put capital spending on the same scale of current expenditures by amortizing the total projects over an assumed bonding period with an underlying common depreciation. They do this in order to compare directly capital spending with current spending in summarizing spending impacts. Because of the constrained nature of capital spending projects, it is generally not possible to compare these expenditures to those of unconstrained increases in budgets. In other words, the impact of funding for the purpose of adding a specific, highly prescribed input is different from the impact of moving the current budget constraint, making it difficult to compare the estimated magnitudes of impact.

Among the studies that use other sources of variation in investments in infrastructure, the nature of the projects varies greatly. Authors study school construction programs in Los Angeles, CA (Lafortune and Schönholzer (2022)), New Haven, CT (Neilson and Zimmerman (2014)), Texas (Schlaffer and Burge (2020)), and England (Zhang, 2014). Conlin and Thompson (2017) also consider state-funded grants for school facility improvement in Ohio. The underlying quality and quantity of capital stock before the implementation of each of these programs varies greatly by context but is not easily measured in a way that facilitates comparison of results across policy contexts.

Using those estimates with medium-term lags of around six years, we find that 13 of the 20 studies report positive effects of capital expenditure on student outcomes. Of these, seven report statistically significant effects. Given the very diverse nature of both the treatment being considered and the specific spending parameter being estimated, it is not possible to provide any reliable quantitative comparison of the results. These results are consistent with the prior spending findings of substantial numbers of estimates that are statistically insignificant and again leading to the conclusion that how funds are used is very important.

Class size

The impact of class size on achievement has been a very controversial policy issue. It was a focus of popular reform in the U.S. after finance policies in California in 1996 included large incentives to reduce class size in grades K-3. That reduction was justified by data from Project STAR, an experimental reduction in class size in Tennessee in the mid-1980s, although as previously described the observational data provided very little support of class size reduction.

Project STAR has received well-justified attention because it was one of the earliest and most policy-relevant RCTs for U.S. school programs. For this experiment, students were randomly assigned to large (23 student) and small (15 student) kindergarten classes in 79 Tennessee schools that volunteered to participate. Students stayed in the original treatment and control groups through grade 3 and were tested at the end of each grade. Unfortunately, by current standards, this experiment suffered from significant biases, and the results have been frequently misinterpreted.[56] Most of the gains found from small classes occurred in kindergarten with much smaller gains in grade 1 and no gains in grades 2 and 3.

The more recent causal evidence on impacts of class size reduction frequently applies one of two popular methodologies: leveraging discrete maximum class size rules (introduced by Angrist and Lavy (1999)) or idiosyncratic population variation (as introduced by Hoxby (2000)) to explore the relationship between class size and student achievement. As seen in Table 15, of the 33 available estimates, most relate to impacts on test scores, and 20 of these test score estimates come from analyses in developed countries outside of the U.S. Interestingly, the evidence from U.S. analyses is

[56] The experiment had large, nonrandom attrition; just 48 percent of the original sample in kindergarten remained in the experiment in grade 3, and those dropping out of the experiment had lower achievement. Entire schools also dropped out of the experiment. There was substantial cross-over from treatment to control group and vice versa. There were significant numbers of missing test scores. There is no information about assignment of teachers to classrooms, and there is no data on the randomization of new students selected to replace students who left the experiment. See Hanushek (1999).

somewhat stronger than that elsewhere with five out of eight U.S. studies showing a statistically significant positive effect (Table 16).

To make the estimates from each study of class size more comparable, we compute the implied effect of a one-student reduction in class size. Because class size is measured at the classroom level and varies between years and sometimes even between class subjects within a student, most estimates are provided for the effect of exposure to smaller class size for one year. Some studies, however, provide estimates of the average effect of exposure to smaller class sizes over a 3-year period (usually 8^{th}-10^{th} grade). We compare the estimates as provided. We scale each estimate by the student-level standard deviation (or baseline averages for attainment and pass rates) and combine estimates across grade levels, populations, and test score subjects using the same method applied for obtaining common parameters in the discussion of school spending studies. It has become somewhat common practice to report effect sizes in terms of 10-student reductions in class size. While this facilitates more transparent comparisons of results across studies, this is not an economically meaningful measure; it is not feasible to engage in reductions of this size, especially in the United States where this would represent a near-halving of class sizes. Thus, we scale results to represent the effects of one-student reductions in class size, assuming that effects are linear in the number of students. The range of test-score results for the U.S. schools, scaled as SD per one student reduction in class size, is shown in Figure 9. The two estimates relying on Project STAR data (Krueger (1999), Krueger and Whitmore (2001)) have much larger estimated impact than the remaining estimates. The median estimate is 0.004 SD improvement per one student reduction with a range from -0.017 to +0.029.

As with the prior estimates about the spending-achievement estimates, the real story, however, is the heterogeneity of the estimates. The estimated percentage of between-study variance is 74 percent.[57] If we discount any possible influence of publication bias or study flaws, we are left with the conclusion that the underlying circumstances that drive the potential impact of class size changes are very important.

Because class size reduction is very expensive, it is useful to compare the magnitude of potential gains with those from spending generally. In the United States, average class size in 2011-12 ranged from 21.2 (elementary grades) to 26.8 (secondary grades).[58] This implies that a 10 percent reduction in class size would, at the median estimate, yield less than 0.01 SD increase in student achievement – an estimated impact dramatically lower than the median estimate of the effect of spending generally in Table 11.[59]

[57] This estimate of the I^2 parameter comes from the eight U.S. studies. If we include all of the non-U.S. studies, the heterogeneity is estimated at 71 percent.

[58] These class sizes compare with 24.1 (elementary) and 23.6 (secondary) in 1993-94. Another perspective is that the overall pupil-teacher ratio in in 2012 was 15.6 (U.S. Department of Education (2020)).

[59] This estimate assumes that total spending changes are consistent with the class size reductions. Smaller class sizes necessitate not only more teachers but also the construction of additional classrooms, the purchase of new classroom materials, additional administrative personnel, and a host of other associated costs.

Salary Policy and Incentives

A natural alternative to the aforementioned input policies is the set of policies based on the performance of teachers.[60] As discussed below, there are many dimensions of such policies, but the unifying theme is adjusting bonuses and salaries of teachers based on student outcomes. Because standard teacher contracts seldom have performance-based components, there are relatively few observational or experimental studies.[61] Moreover, this is an area where the majority of modern empirical studies consider experiments outside of the United States.

A fundamental problem in the evaluation of teacher incentive programs is, nonetheless, that they focus on the effort margin and ignore the selection margin. Specifically, they evaluate the impact of changed incentives on the existing stock of teachers and consider whether they perform better after the introduction of specific monetary incentives – the effort margin. They do not consider whether different incentive schemes lead through entry and exit of teachers to a difference quality distribution of teachers – the selection margin.[62] Investigating the effort margin makes more sense in the case of many developing countries where rampant teacher absences provide a natural focus for incentives based on effort, but they have less relevance in the U.S. and other developed countries where teacher absences are low and where teachers generally are focused on student learning.[63]

The incentive designs used in performance pay take on a variety of forms, and it is unclear which combination of these aspects will be most effective in inducing achievement gains. Individual incentives reward teachers for performance of only their own class over the year, though there may be negative impacts on the collaborative nature of teachers under this scheme that also contribute to its political infeasibility. Group incentives reward all teachers in each grade level or school for aggregate performance, inducing collaboration among teachers while potentially introducing free-riding. In tournaments, teachers or groups of teachers are ranked based upon their students' performance on an

[60] Related, it is possible to think of policies involving school performance. Such policies, which generally fall under the heading of school accountability, are not considered here because they typically are not thought of as resource policies; see Hanushek and Raymond (2005), Dee and Jacob (2011).

[61] There are major exceptions to standard teacher contracts in Washington, DC, and in Dallas, TX. Evaluations of these show substantial impacts on student performance (see Dee and Wyckoff (2015),2017)).

[62] An exception is found in the developing country literature. Brown and Andrabi (2020) design an experiment in Pakistani schools that allows for both the assessment of teacher choices to select into roles with performance-related pay over fixed wages and the analysis of the effect of assignment to a performance-related pay scheme. They find that teachers with higher ability and responsiveness to effort incentives tend to select into performance-related schemes, where they can expect to earn higher than the base fixed wage given their expectations of their own performance. Critically, estimates that account for these sorting effects are twice as large as those that only consider the effort margin, suggesting that it will be vital for further research to investigation selection effects in all contexts.

[63] For example, an RCT in Indian schools tied financial incentives to teacher attendance instead of student performance (Duflo, Hanna, and Ryan (2012)). Salaries for teachers treated in this experiment were made a non-linear function of attendance each month, and absenteeism improved by 21 percentage points relative to the fixed wage teachers, which in turn led to a 0.17 standard deviation improvement in student scores. At baseline, teachers in program schools had an absence rate of 44%. In 2013, the average absence rate of public school teachers in the U.S. was less than 10% (Saenz-Armstrong (2020)). Most teachers are in school nearly every day. On the other hand, some research has also pointed to the importance of considering teacher absences in the U.S. (Hansen and Quintero (2020)).

achievement metric, and a selection of the highest-ranking teachers will receive bonuses. There also may be a sliding scale of bonuses based upon ranking. In piece-rate incentive structures, all teachers that score above a threshold on the achievement metric receive a bonus commensurate with that threshold. Metrics of achievement vary across programs as well, with some programs opting to reward teachers for average scores (levels) while others reward improvements in student scores from year to year (gains). Finally, incentive schemes differ in the size of the payments, with average additional payments tied to student performance ranging experimentally from 2-15 percent of base annual teacher pay and bonuses tied to teacher attendance reaching up to 30 percent of base salary.

A few examples provide a perspective on both the approaches to evaluation and the results. Fryer (2013) evaluates the effectiveness of teacher incentive pay through a randomized controlled trial in New York City elementary schools with high poverty rates. The program assigned treatment schools to a group-based incentive scheme in which a score made up of a combination of improvement in exam proficiency, performance as measured by exam pass rates and graduation, and a measure of school environment including attendance rates determined teacher bonuses at the school level. An instrumental variables estimation using program assignment as an instrument for program participation suggests that there were no positive effects of group-based teacher bonus incentive schemes on student achievement in grades 3-8. Fryer posits that these null effects can be attributed to flaws in the incentive design that prevented teachers from being able to transparently predict how their efforts would translate into rewards. Goodman and Turner (2013) and Marsh et al. (2011) similarly find that the New York City bonus program for high poverty public schools had no effect on student achievement. Goodman and Turner note that some larger incentives had small positive effects on teacher effort, suggesting that the structure and size of the bonus payments may not have been properly calibrated to elicit additional teacher effort.

The Project on Incentives in Teaching (POINT) was a three-year study conducted in the Metropolitan Nashville School System from 2006-07 through 2008-09 (Springer et al. (2010)). Middle school math teachers voluntarily participated in a controlled experiment to assess the effect of financial rewards for teachers whose students showed unusually large gains on standardized tests. Students of teachers randomly assigned to the treatment group (eligible for bonuses) did not outperform students whose teachers were assigned to the control group (not eligible for bonuses). However, attrition of teachers from POINT was high with half of the initial participants leaving before the end of the experiment. Thus, differential selection of exiting teachers may have influenced the results.

Though the evidence of the effectiveness of performance pay in the United States is limited, this is one of the richest areas of evidence on the role of resources in education in developing countries. For example, field experiments in Tanzania (Mbiti et al. (2019)), India (Muralidharan and Sundararaman (2011)), Kenya (Glewwe, Ilias, and Kremer (2010)), and Guinea (Barrera-Osorio, Cilliers, Cloutier, and Filmer (2022)) yielded positive results, albeit of different magnitudes.

The available studies are shown in Table 17. It is difficult to construct comparable impact parameters because only 21 of the 31 estimates provide information to construct comparable measures of intensity of the salary incentives. For studies in which the information regarding the value of payment is available, average additional payments tied to student performance range from 2-15 percent

of base annual teacher pay. Because of the missing information, however, we present the estimated effect of switching from a traditional payment scheme to some variation of an incentive pay scheme without regard for intensity of the incentive. We scale each estimate by the student-level standard deviation (or baseline averages for attainment and pass rates) and combine estimates across grade levels, populations, and test score subjects using the same method applied for obtaining common parameters in the discussion of school spending studies.

Table 18 presents summary data on the available incentive evaluations. We find that the median estimate of all studies implies that a switch to performance-related pay yields a .074 standard deviation increase in student achievement. For U.S. studies, this value is 0.048 standard deviations in achievement, but as shown in Figure 10 the estimates range from -0.20 to 0.158, and only 4 of the 11 estimates are statistically significant.[64] The U.S. results again show wide variation in the impact of incentives directed at the effort margin, but there is a complete lack of information about the important issue of the selection margin. The results do imply policy uncertainty when just the effort margin is considered.

6 Some Open questions

A range of follow-on questions have been exposed by this review of recent evidence. One of the largest is the need for replication of the results. The studies included here focus on identification of causal impacts of resources on outcomes. Under increasingly well-understood conditions, the various methods provide heightened internal validity of the estimation that leads to unbiased estimates of impact parameters. But an unbiased estimator does not ensure that any single estimate will be close to the true impact. Nor does the set of results in school finance indicate clearly the circumstances under which they can be generalized to other programs and policies.

It is difficult in the case of the resource-outcome estimates presented here to know exactly how to replicate the analyses.[65] In addition to the normal incentives against replication,[66] studies in this area face a particular design difficulty. It is poorly understood the extent that the estimated impact parameters are sensitive to the restrictions on specific spending, that key parts of the institutional structure of the state educational systems are important, and that the particular subset and cohort of the students enters into the response.

The importance of understanding why resources appear to have much larger impacts in some situations rather than in others is critical (assuming that the estimated differences are more than either sampling error, flaws in the underlying studies, or some form of publication bias). In other words, how

[64] These results mirror the historic findings of Cohen and Murnane (1985, 1986). They found that merit pay had little impact because the added bonus was both small and transitory.
[65] The term replication is used loosely here and includes robustness analyses in the definition of Clemens (2017).
[66] The incentives against replication (for both authors and editors) are the backdrop of the long-discussed issues around publication bias. See section 5.1, above. The statistical demands on replication are also large unless the original studies have unusually high power (Hedges and Schauer (2019)).

resources are used appears to be key, but we are currently lacking any general rules from which it is possible to interpret the existing resource-achievement estimates – or, importantly, to use the aggregate evidence in policy decisions.

Another issue that has received considerable parallel attention but that has not entered centrally into the analyses of resources and outcomes is the measurement of outcomes. The general analysis involves use of proxies for labor market outcomes and not the outcomes themselves.[67] The studies also stop short of considering alternatives to the specific outcome measures that are available. A range of studies has emphasized various noncognitive measures of student outcomes in studying general educational production.[68] But, even in the cognitive range, there are questions about differences across domains as evident by the varying patterns of score changes between reading and math.[69] If there are multiple outcomes, they tend to be treated simply as alternatives, and the analysis proceeds ad seriatim. These variations across outcomes have never been fully investigated in the context of resources and schools.

There has been considerable policy discussion about the importance of pre-school education.[70] Because most states separate funding of preschool from funding for K-12 education, the previous discussions of both costs and outcomes do not delve into any consideration of the interactions with preschool programs. Because early childhood programs and K-12 programs are complementary, it makes sense to consider how spending on each fits together to impact outcomes. In particular, since there is a trade-off in where public funds go, it would be very valuable to consider how the current funding patterns might be altered to improve student outcomes.

Finally, it is not difficult to attribute any inefficiencies in spending and operation of the schools to constraints on the system. Schools do not operate in unregulated markets but instead are subject to substantial regulations covering everything from hiring rules to operational details about class size or length of the school day. The implications of such regulations are poorly understood but almost certainly are part of the picture of spending-achievement relationships and why results are apparently so different across states.

Finally, one particular constraint that has surprisingly received very limited analysis is the role of teachers' unions on the operations of schools. A limited number of studies have investigated the impact on cost and outcomes of schools, suggesting that unions do affect schools (Lovenheim (2009), Lovenheim and Willén (2019), Moe (2011)), but such studies are remarkably few and limited compared to the pervasiveness of hypothesized influence. It appears very likely that restrictions from unionized bargaining and contracts interact significantly with resource decisions.

---

[67] An exception is Jackson, Johnson, and Persico (2016), which considers labor market and other outcomes.
[68] See, for example, Heckman, Stixrud, and Urzua (2006), Cunha and Heckman (2008), Borghans, Duckworth, Heckman, and Weel (2008), Mendez (2015), West et al. (2016).
[69] A very common finding of research into educational production functions and into teacher quality is that schools and teachers have a greater impact on math than on reading (e.g., Hanushek and Rivkin (2010, 2012)). Relatedly, relative skill differences are emphasized in Hanushek et al. (2021).
[70] See, for example, Barnett (1992), Belfield, Nores, Barnett, and Schweinhart (2006), Finn (2009), Havnes and Mogstad (2015), Heckman et al. (2010), Whitehurst (2018),

7 Conclusions

The recent rapid expansion of studies delving into the relationship of resources and outcomes has added considerably to understanding what is possible from various educational decisions. Most importantly, the recent studies have applied the arsenal of empirical techniques designed to probe causality to the crucial questions of how to improve educational outcomes. Importantly, these newest studies have reinforced the prior conclusion that how money and resources are applied is crucial to the results.

The United States has a long history of trying to improve the achievement and skills of its students, particularly of its disadvantaged students. Beginning with the "War on Poverty" that commenced in the 1960s, the U.S. has expanded funding of students. This expansion has been led by the separate states and localities, since educational decision making is largely the province of the individual states. But, unique to the United States, the state courts have played a very active role in decisions about school finance. In a multitude of decisions, separate state courts have entered into discussions of equity and of adequacy of funding. Because of the limited role of the courts, however, judicial decisions are generally restricted to the distribution of funds.

The result of the combination of legislative and court decisions has been a significant expansion of funding for schools. Evidence points to a modest closing of achievement gaps between advantaged and disadvantaged, but the slow pace of closure implies that significant inequality will persist for a very long time. The record in terms of the level of performance is mixed, with some evidence of improvement at earlier grades but little evidence of improvement at later grades. Moreover, improvements have been concentrated in early-grade math performance. (The pandemic experiences, however, erased much of the prior improvement).

The historic empirical research showed limited relationships between standard measures of school resources and student outcomes. It was, however, rightfully questioned because of concerns about the quality of many studies and especially about the potential for biases from omitted variables and endogeneity of the measured resources. The investigations of educational production functions did, nonetheless, introduce credible evidence about the heterogeneity of resource effects and also introduced questions about overall inefficiency of resource decisions.

The "credibility revolution" of modern empirical economic analysis has deeply penetrated recent analysis of educational resources and outcomes. These explorations exploit exogenous variation in resources from a variety of sources to consider how funding of schools impacts student outcomes as measured by test scores, test passing rates, or continuation in schooling. We have attempted to compile the results of all high quality analyses that provide direct evidence on the impact of added resources. This search includes both published and unpublished studies and analyses from around the world, although our main emphasis is studies of U.S. schools. We have taken the estimates as produced and have for the most part ignored any possible influence of flawed analytics or of publication bias.

It is difficult to make direct comparisons of the results across all of the studies, but the analyses of test scores – arguably the most important of the measures – can be most readily linked and assessed. The existing spending parameter estimates are all transformed into a common metric, the achievement impact measured in terms of individual student standard deviations of a 10 percent increase in spending. The 16 studies of spending for the U.S. have a median effect size of 0.07 SD per 10 percent

increase, but the study estimates range from -0.244 SD to +0.543. While most point estimates are positive, only nine are statistically different from zero at the 5 percent level.

The wide variation in estimated effects can be partially attributed to sampling errors, but a larger element is likely to be systematic differences in the precise spending parameter that is being estimated. These estimates are derived from observations of spending changes under very different settings. They range from dramatic changes of the funding formula within a single state to recession-induced spending reductions to legislative responses to legal judgments across multiple states to differences in federal compensatory aid for disadvantaged students. As such, the estimated spending impacts each apply to specific circumstances.

The methodologies are designed around ideas of internal validity that produce unbiased estimates of the impact of spending increases, but the impacts are not independent of the conditions that govern the use and effect of the added resources. Thus, for example, knowledge of the effect on student achievement of disadvantaged children from added compensatory funds through the federal Title I program does not necessarily provide direct information about spending choices under unrestricted movements in the school budget constraint.

The median estimate of spending impacts does not easily generalize to the historical movements in school spending for the U.S. as a whole. From performance data going back to 1978, it is possible to trace the change in achievement and the corresponding aggregate spending data for various age cohorts in reading and math performance. The median estimated impact parameter from the 16 spending studies is less than the unconditional spending-achievement changes observed for mathematics in lower age/grade groups but significantly exceeds that for all reading performance measures and for math at age 17.[71]

Of course, the unconditional spending-achievement changes reflect a combination of school impacts and other impacts such as family, peers, and neighborhood. But the differences between the median impact parameter and the unconditional historical data is very large, implying wide swings in the non-school component would be needed in order to reconcile the impact parameters with the national data.

This new evidence on spending impacts, like the historical evidence, does not indicate that spending does not matter. Nor does it indicate that spending cannot matter. It does indicate that simply adding more resources without addressing how and where the resources will be used provides little assurance that student achievement will improve. Little progress has been made leveraging the results to uncover when more spending will have significant impacts and when it will not.

The spending impact on school attainment, while more consistent across studies, is harder to interpret. It is more difficult to interpret because attainment ignores quality differences and is dramatically affected by individual behavioral responses to differences in costs and returns of further schooling. The consistency comes from finding more statistically significant positive impacts of spending, but again, these estimated impacts vary widely across studies and are difficult to reconcile with the historical data. The learning losses during the pandemic make interpretations of these changes

[71] These findings relate to pre-pandemic outcomes. Achievement fell sharply after school closures in March 2020 and during the subsequent school years. See Section 2.3.

particularly challenging because they show dramatically the significant differences in achievement associated with differences in school attainment.

The analyses of specific input changes offer an additional picture of how policies directed at classes of inputs affect student outcomes. The heterogeneity of these input studies reinforces the message that how resources should be used goes beyond simple input mechanisms including capital investments, class size reduction, and teacher incentives. As with the simple spending impact studies, there is a range of estimates – many of which are very imprecisely estimated – that come out of the available contemporaneous studies, but there is no clear description of when (or if) instituting such policies is efficacious.

References

Abdulkadiroğlu, Atila, and Tommy Andersson. 2023. "School choice." In *Handbook of the Economics of Education, Vol 6*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann: Elsevier: 135-185.

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

Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *The Quarterly Journal of Economics* 114, no. 2 (1999/05/01/): 533-575.

Angrist, Joshua, Peter Hull, and Christopher R. Walters. forthcoming 2023. "Methods for Measuring School Effectiveness." In *Handbook on the Economics of Education, Vol 7*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann: Elsevier.

Armor, David J., Patricia Conry-Oseguera, Millicent Cox, Niceima King, Lorraine McDonnell, Anthony Pascal, Edward Pauly, and Gail Zellman. 1976. *Analysis of the school preferred reading program in selected Los Angeles minority schools*. Santa Monica, CA: Rand Corp.

Bacher-Hicks, Andrew, Mark J. Chin, Thomas J. Kane, and Douglas O. Staiger. 2017. "An Evaluation of Bias in Three Measures of Teacher Quality: Value-Added, Classroom Observations, and Student Surveys." NBER Working Paper No. 23478. Cambridge, MA: National Bureau of Economic Research (June).

Bacher-Hicks, Andrew, and Cory Koedel. 2023. "Estimation and interpretation of teacher value added in research applications." In *Handbook of the Economics of Education, Vol 6*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann: Elsevier: 93-134.

Barnett, W. Steven. 1992. "Benefits of compensatory preschool education." *Journal of Human Resources* 27, no. 2 (Spring): 279-312.

Baron, E. Jason, Joshua M. Hyman, and Brittany N. Vasquez. 2022. Public School Funding, School Quality, and Adult Crime: National Bureau of Economic Research.

Barrera-Osorio, Felipe, Jacobus Cilliers, Marie-Hélène Cloutier, and Deon Filmer. 2022. "Heterogenous teacher effects of two incentive schemes: Evidence from a low-income country." *Journal of Development Economics* 156(May): 102820.

Belfield, Clive R., Milagros Nores, Steve W. Barnett, and Lawrence J. Schweinhart. 2006. "The High/Scope Perry Preschool Program." *Journal of Human Resources* 41, no. 1 (Winter): 162-190.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2021. *Introduction to Meta-Analysis*. Second ed. London: John Wiley & Sons.

Borghans, Lex, Angela Lee Duckworth, James J. Heckman, and Bas ter Weel. 2008. "The Economics and Psychology of Personality Traits." *Journal of Human Resources* 43, no. 4 (October 2, 2008): 972-1059.

Bound, John, Michael F. Lovenheim, and Sarah Turner. 2010. "Why have college completion rates declined? An analysis of changing student preparation and collegiate resources." *American Economic Journal: Applied Economics* 2, no. 3 (July): 129-157.

Bowles, Samuel, and Henry M. Levin. 1968. "The determinants of scholastic achievement--an appraisal of some recent evidence." *Journal of Human Resources* 3, no. 1 (Winter): 3-24.

Boyd, Donald, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2011. "Teacher Layoffs: An Empirical Illustration of Seniority versus Measures of Effectiveness." *Education Finance and Policy* 6, no. 3 (Summer): 439-454.

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

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

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

Brown, Christina, and Tahir Andrabi. 2020. "Inducing positive sorting through performance pay: Experimental evidence from Pakistani schools." *University of California at Berkeley Working Paper*.

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

Buerger, Christian, Seung Hyeong Lee, and John D. Singleton. 2021. "Test-Based Accountability and the Effectiveness of School Finance Reforms." *AEA Papers and Proceedings* 111: 455-459.

Burtless, Gary, ed. 1996. *Does money matter? The effect of school resources on student achievement and adult success*. Washington, DC: Brookings.

Cain, Glen G., and Harold W. Watts. 1970. "Problems in making policy inferences from the Coleman Report." *American Sociological Review* 35, no. 2 (April): 328-352.

Candelaria, Christopher A., and Kenneth A. Shores. 2019. "Court-Ordered Finance Reforms in the Adequacy Era: Heterogeneous Causal Effects and Sensitivity." *Education Finance and Policy* 14, no. 1: 31-60.

Cascio, Elizabeth U., Nora Gordon, and Sarah Reber. 2013. "Local Responses to Federal Grants: Evidence from the Introduction of Title I in the South." *American Economic Journal: Economic Policy* 5, no. 3: 126-159.

Cellini, Stephanie Riegg, Fernando Ferreira, and Jesse Rothstein. 2008. *The Value of School Facilities: Evidence from a Dynamic Regression Discontinuity Design*.

Chetty, Raj, John N. Friedman, and Jonah Rockoff. 2014. "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood." *American Economic Review* 104, no. 9 (September): 2633-2679.

Clemens, Michael A. 2017. "The meaning of failed replications: A review and proposal." *Journal of Economic Surveys* 31, no. 1: 326-342.

Cohen, David K., and Richard J. Murnane. 1985. "The merits of merit pay." *Public Interest* 80(Summer): 3-30.

Cohen, David K., and Richard J. Murnane. 1986. "Merit pay and the evaluation problem: Understanding why most merit pay plans fail and a few survive." *Harvard Educational Review* 56, no. 1 (February): 1-17.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York. 1966. *Equality of educational opportunity*. Washington, D.C.: U.S. Government Printing Office.

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

Coons, John E., William H. Clune, and Stephen D. Sugarman. 1970. *Private wealth and public education*. Cambridge, MA: The Belknap Press of Harvard University Press.

Cornman, Stephen Q., J.J. Phillips, Malia R. Howell, and Lei Zhou. 2022. *Revenues and Expenditures for Public Elementary and Secondary Education: FY20*, NCES 2022-301. Washington, DC: National Center for Education Statistics (May).

CREDO. forthcoming 2023. *National Charter School Study III*. Stanford University: Center for Research on Educational Outcomes.
Cunha, Flavio, and James J. Heckman. 2008. "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *The Journal of Human Resources* 43, no. 4 (Fall): 738-782.
Dee, Thomas S., and Brian A. Jacob. 2011. "The impact of No Child Left Behind on student achievement." *Journal of Policy Analysis and Management* 30, no. 3: 418-446.
Dee, Thomas S., and James Wyckoff. 2015. "Incentives, Selection, and Teacher Performance: Evidence from IMPACT." *Journal of Policy Analysis and Management* 34, no. 2: 267-297.
Dee, Thomas S., and James Wyckoff. 2017. "A Lasting Impact: High-stakes teacher evaluations drive student success in Washington, D.C." *Education Next* 17, no. 4 (Fall): 58-66.
Dills, Angela K., Patrick J. Wolf, Corey A. DeAngelis, Jay F. May, Larry D. Maloney, and Cassidy Syftestad. 2021. *Charter School Funding: Dispelling Myths about EMOs, Expenditure Patterns, & Nonpublic Dollars*. School Choice Demonstration Project. Fayetteville, AR: Department of Education Reform, University of Arkansas (October).
Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives work: Getting teachers to come to school." *American Economic Review* 102, no. 4: 1241-1278.
Dynarski, Susan, Aizat Nurshatayeva, Lindsay C. Page, and Judith Scott-Clayton. 2023. "Addressing nonfinancial barriers to college access and success: Evidence and policy implications." In *Handbook of the Economics of Education*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann: Elsevier.
Dynarski, Susan, Lindsay Page, and Judith Scott-Clayton. 2023. "Financial Aid for College Students." In *Handbook of the Economics of Education, Vol. 7*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann: Elsevier.
Finn, Chester E., Jr. 2009. *Reroute the pre-school juggernaut*. Stanford, CA: Hoover Institution Press.
Fischel, William A. 1989. "Did Serrano cause Proposition 13?" *National Tax Journal* 42(December): 465-474.
Fischel, William A. 2006. "The courts and public school finance: Judge-made centralization and economic research." In *Handbook of the Economics of Education*, edited by Eric A. Hanushek and Finis Welch. Amsterdam: North Holland: 1277-1325.
Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science* 345, no. 6203 (August 28): 1502-1505.
Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher incentives." *American Economic Journal: Applied Economics* 2, no. 3 (July): 205-227.
Goldhaber, Dan, and Michael Hansen. 2013. "Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance." *Economica* 80, no. 319 (July): 589-612.
Goldhaber, Dan, and Roddy Theobald. 2013. "Managing the Teacher Workforce in Austere Times: The Determinants and Implications of Teacher Layoffs." *Education Finance and Policy* 8, no. 4 (Fall): 494-527.
Goldhaber, Dan; Hannaway, Jane. 2009. *Creating a New Teaching Profession*. Washington, D.C.: The Urban Institute Press.
Goldstein, Jessica, and Josh B. McGee. 2020. "Did Spending Cuts During the Great Recession Really Cause Student Outcomes to Decline?" EdWorkingPaper No. 20-303. Brown University: Annenberg (October).
Goncalves, Felipe. 2015. "The Effects of School Construction on Student and District Outcomes: Evidence from a State-Funded Program in Ohio." (2015-11-05).

Goodman, Sarena F., and Lesley J. Turner. 2013. "The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program." *Journal of Labor Economics* 31, no. 2 (April): 409-420.
Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. "Identifying effective teachers using performance on the job." Hamilton Project Washington: Brookings Institution (April).
Guryan, Jonathan. 2001. Does Money Matter? Regression-Discontinuity Estimates from Education Finance Reform in Massachusetts: National Bureau of Economic Research.
Halloran, Clare, Rebecca Jack, James C. Okun, and Emily Oster. 2021. "Pandemic Schooling Mode and Student Test Scores: Evidence from US States." *National Bureau of Economic Research Working Paper Series* No. 29497.
Hansen, Michael, and DIana Quintero. 2020. We should be focusing on absenteeism among teachers, not just students. In *Brown Center Chalkboard*. Washington, DC: Brookings Institution.
Hanushek, Eric A, and Matthew Wirtz. forthcoming. "A portrait of court involvement in school finance." Stanford University: Hoover Institution.
Hanushek, Eric A. 1971. "Teacher characteristics and gains in student achievement: Estimation using micro data." *American Economic Review* 60, no. 2 (May): 280-288.
Hanushek, Eric A. 1997. "Assessing the effects of school resources on student performance: An update." *Educational Evaluation and Policy Analysis* 19, no. 2 (Summer): 141-164.
Hanushek, Eric A. 1999. "Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects." *Educational Evaluation and Policy Analysis* 21, no. 2 (Summer): 143-163.
Hanushek, Eric A. 2002. "Publicly provided education." In *Handbook of Public Economics, Vol. 4*, edited by Alan J. Auerbach and Martin Feldstein. Amsterdam: North Holland: 2045-2141.
Hanushek, Eric A. 2003. "The failure of input-based schooling policies." *Economic Journal* 113, no. 485 (February): F64-F98.
Hanushek, Eric A. 2011. "The economic value of higher teacher quality." *Economics of Education Review* 30, no. 3 (June): 466-479.
Hanushek, Eric A., Babs Jacobs, Guido Schwerdt, Rolf van der Velden, Stan Vermeulen, and Simon Wiederhold. 2021. "The Intergenerational Transmission of Cognitive Skills: An Investigation of the Causal Impact of Families on Student Outcomes." NBER Working Paper No. 29450. Cambridge, MA: National Bureau of Economic Research (December).
Hanushek, Eric A., and John F. Kain. 1972. "On the value of `equality of educational opportunity' as a guide to public policy." In *On equality of educational opportunity*, edited by Frederick Mosteller and Daniel P. Moynihan. New York: Random House: 116-145.
Hanushek, Eric A., Jacob D. Light, Paul E. Peterson, Laura M. Talpey, and Ludger Woessmann. 2022a. "Long-run Trends in the U.S. SES-Achievement Gap." *Education Finance and Policy* 17, no. 4: 608-640.
Hanushek, Eric A., Jacob Light, Paul E. Peterson, Laura M. Talpey, and Ludger Woessmann. 2022b. "Long-run Trends in the U.S. SES-Achievement Gap." *Education Finance and Policy* 17, no. 4 (Fall): 608-640.
Hanushek, Eric A., and Alfred A. Lindseth. 2009. *Schoolhouses, courthouses, and statehouses: Solving the funding-achievement puzzle in America's public schools*. Princeton, NJ: Princeton University Press.
Hanushek, Eric A., and Margaret E. Raymond. 2005. "Does school accountability lead to improved student performance?" *Journal of Policy Analysis and Management* 24, no. 2: 297-327.
Hanushek, Eric A., and Steven G. Rivkin. 1997. "Understanding the twentieth-century growth in U.S. school spending." *Journal of Human Resources* 32, no. 1 (Winter): 35-68.

Hanushek, Eric A., and Steven G. Rivkin. 2010. "Generalizations about using value-added measures of teacher quality." *American Economic Review* 100, no. 2 (May): 267-271.
Hanushek, Eric A., and Steven G. Rivkin. 2012. "The distribution of teacher quality and implications for policy." *Annual Review of Economics* 4: 131-157.
Hanushek, Eric A., Steven G. Rivkin, and Lori L. Taylor. 1996. "Aggregation and the estimated effects of school resources." *Review of Economics and Statistics* 78, no. 4 (November): 611-627.
Hanushek, Eric A., Guido Schwerdt, Simon Wiederhold, and Ludger Woessmann. 2015. "Returns to skills around the world: Evidence from PIAAC." *European Economic Review* 73: 103-130.
Hanushek, Eric A., Guido Schwerdt, Simon Wiederhold, and Ludger Woessmann. 2017. "Coping with change: International differences in the returns to skills." *Economics Letters* 153: 15-19.
Hanushek, Eric A., and Ludger Woessmann. 2015. *The knowledge capital of nations: Education and the economics of growth*. Cambridge, MA: MIT Press.
Hanushek, Eric A., and Ludger Woessmann. 2020. *The Economic Impacts of Learning Losses*. Paris: OECD (September).
Harris, Douglas N. 2011. *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press.
Havnes, Tarjei, and Magne Mogstad. 2015. "Is universal child care leveling the playing field?" *Journal of Public Economics* 127(2015/07/01/): 100-114.
Head , Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. "The Extent and Consequences of P-Hacking in Science." *PlosBiology* 13, no. 3 (March 13).
Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz. 2010. "The rate of return to the HighScope Perry Preschool Program." *Journal of Public Economics* 94, no. 1–2 (February): 114-128.
Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior." *Journal of Labor Economics* 24, no. 3 (July): 411-482.
Hedges, Larry V., Richard D. Laine, and Rob Greenwald. 1994. "Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes." *Educational Researcher* 23, no. 3 (April): 5-14.
Hedges, Larry V., and Jacob M. Schauer. 2019. "More Than One Replication Study Is Needed for Unambiguous Tests of Replication." *Journal of Educational and Behavioral Statistics* 44, no. 5: 543–570.
Ho, Andrew Dean. 2008. "The Problem With "Proficiency": Limitations of Statistics and Policy Under No Child Left Behind." *Educational Researcher* 37, no. 6 (August 1, 2008): 351-360.
Holland, Paul W. 2002. "Two Measures of Change in the Gaps Between the CDFs of Test-Score Distributions." *Journal of Educational and Behavioral Statistics* 27, no. 1: 3-17.
Hoxby, Caroline M. 2000. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *The Quarterly Journal of Economics* 115, no. 4 (2000/11/01/): 1239-1285.
Hyman, Joshua. 2017. "Does Money Matter in the Long Run? Effects of School Spending on Educational Attainment." *American Economic Journal: Economic Policy* 9, no. 4: 256-280.
Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos. 2017. "The Power of Bias in Economics Research." *The Economic Journal* 127, no. 605 (October): F236-F265.
Jackson, C. Kirabo, Rucker C. Johnson, and Claudia Persico. 2016. "The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms." *Quarterly Journal of Economics* 131, no. 1 (February): 157-218.
Jackson, C. Kirabo, and Claire Mackevicius. 2021. "The Distribution of School Spending Impacts." NBER Working Paper No. 28517. Cambridge, MA: National Bureau of Economic Research (March).

Jackson, C. Kirabo, Jonah E. Rockoff, and Douglas O. Staiger. 2014. "Teacher effects and teacher related policies." *Annual Review of Economics* 6: 801-825.
Jackson, C. Kirabo, Cora Wigger, and Heyu Xiong. 2021. "Do School Spending Cuts Matter? Evidence from the Great Recession." *American Economic Journal: Economic Policy* 13, no. 2 (May): 304-335.
Johnson, Rucker C. 2015. "Follow the Money: School Spending from Title I to Adult Earnings." *RSF: The Russell Sage Foundation Journal of the Social Sciences* 1, no. 3 (December): 50-76.
Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff. 2015. "Value-added modeling: A review." *Economics of Education Review* 47: 180-195.
Kranz, Sebastian, and Peter Pütz. 2022. "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics: Comment." *American Economic Review* 112, no. 9 (September): 3124-36.
Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *The Quarterly Journal of Economics* 114, no. 2 (1999): 497-532.
Krueger, Alan B., and Diane M. Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR." *The Economic Journal* 111, no. 468 (January): 1-28.
Kuhfeld, Megan, James Soland, and Karyn Lewis. 2022. "Test Score Patterns Across Three COVID-19-impacted School Years." EdWorkingPapers 22-521. Brown University: Annenberg.
Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach. 2018. "School finance reform and the distribution of student achievement." *American Economic Journal: Applied Economics* 10, no. 2: 1-26.
Lafortune, Julien, and David Schönholzer. 2022. "The Impact of School Facility Investments on Students and Homeowners: Evidence from Los Angeles." *American Economic Journal: Applied Economics* 14, no. 3: 254-89.
Lee, Kyung-Gon, and Solomon W. Polachek. 2018. "Do school budgets matter? The effect of budget referenda on student dropout rates." *Education Economics* 26, no. 2 (March): 129-144.
Lovenheim, Michael F., and Alexander Willén. 2019. "The Long-Run Effects of Teacher Collective Bargaining." *American Economic Journal: Economic Policy* 11, no. 3 (August): 292-324.
Lovenheim, Michael F. 2009. "The effect of teachers' unions on education production: Evidence from union election certifications in three Midwestern states." *Journal of Labor Economics* 27, no. 4 (October): 525-587.
Marsh, Julie A., Matthew G. Springer, Daniel F. McCaffrey, Kun Yuan, Scott Epstein, Julia Koppich, Nidhi Kalra, Catherine DiMartino, and Art Peng. 2011. *A Big Apple for Educators: New York City's Experiment with Schoolwide Performance Bonuses: Final Evaluation Report*: RAND Corporation (2011/07/18/).
Martorell, Paco, Kevin Stange, and Isaac McFarlin. 2016. "Investing in schools: capital spending, facility conditions, and student achievement." *Journal of Public Economics* 140(August): 13-29.
Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2019. "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania." *The Quarterly Journal of Economics* 134, no. 3: 1627-1673.
Mendez, Ildefonso. 2015. "The effect of the intergenerational transmission of noncognitive skills on student performance." *Economics of Education Review* 46(2015/06/01/): 78-97.
Miller, Corbin L. 2018. "The Effect of Education Spending on Student Achievement: Evidence from Property Values and School Finance Rules." 122.
Moe, Terry M. 2011. *Special interest: Teachers unions and America's public schools*. Washington, DC: Brookings Institution Press.
Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher performance pay: Experimental evidence from India." *Journal of Political Economy* 119, no. 1 (February): 39-77.

Murnane, Richard J. 1975. *Impact of school resources on the learning of inner city children*. Cambridge, MA: Ballinger.

Neilson, Christopher A., and Seth D. Zimmerman. 2014. "The effect of school construction on test scores, school enrollment, and home prices." *Journal of Public Economics* 120(December): 18-31.

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

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

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

Panhans, Matthew T., and John D. Singleton. 2017. "The Empirical Economist's Toolkit: From Models to Methods." *History of Political Economy* 49, no. Supplement: 127-157.

Papke, Leslie E. 2008. "The Effects of Changes in Michigan's School Finance System." *Public Finance Review* 36, no. 4 (July): 456-474.

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

Reardon, Sean F. 2011. The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances*, edited by Richard J. Murnane and Greg J. Duncan. New York: Russell Sage Foundation: 91-116.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, schools, and academic achievement." *Econometrica* 73, no. 2 (March): 417-458.

Rothstein, Jesse. 2015. "Teacher Quality Policy When Supply Matters." *American Economic Review* 105, no. 1: 100-130.

Roy, Joydeep. 2011. "Impact of School Finance Reform on Resource Equalization and Academic Performance: Evidence from Michigan." *Education Finance and Policy* 6, no. 2 (Spring): 137-167.

Saenz-Armstrong, Patricia. 2020. *Roll Call 2020*. Washington, DC: NCTQ (December).

Schlaffer, James, and Gregory Burge. 2020. "The asymmetric effects of school facilities on academic achievement: Evidence from Texas bond votes." *The Social Science Journal*(April): 1-19.

Silva, Fabio, and Jon Sonstelie. 1995. "Did Serrano cause a decline in school spending?" *National Tax Journal* 48, no. 2 (June): 199-215.

Springer, Matthew G., Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J.R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher. 2010. *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching*. Nashville, TN: National Center on Performance Incentives, Vanderbilt University.

Staiger, Douglas O., and Jonah E. Rockoff. 2010. "Searching for Effective Teachers with Imperfect Information." *Journal of Economic Perspectives* 24, no. 3 (Summer): 97–118.

Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the specification and estimation of the production function for cognitive achievement." *Economic Journal* 113, no. 485: F3-33.

U.S. Department of Education. 2020. *Digest of Education Statistics 2020*. Washington, DC: National Center for Education Statistics.

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

West, Martin R., Matthew A. Kraft, Amy S. Finn, Rebecca E. Martin, Angela L. Duckworth, Christopher F. O. Gabrieli, and John D. E. Gabrieli. 2016. "Promise and Paradox: Measuring Students' Non-Cognitive Skills and the Impact of Schooling." *Educational Evaluation and Policy Analysis* 38, no. 1 (March): 148-170.

Whitehurst, Grover J. 2018. *Does state pre-K improve children's achievement?* Evidence Speaks Reports. Washington, DC: Brookings Institution (July 12).
Wise, Arthur E. 1968. *Rich Schools, Poor Schools; the Promise of Equal Educational Opportunity*. Chicago: University of Chicago Press.

Studies Analyzed (Tables 10, 14, 15, and 17)

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

Andrabi, Tahir, and Christina Brown. 2022. "Subjective versus Objective Incentives and Employee Productivity." (mimeo) (February 22).

Angrist, Joshua D., Erich Battistin, and Daniela Vuri. 2017. "In a Small Moment: Class Size and Moral Hazard in the Italian Mezzogiorno." *American Economic Journal: Applied Economics* 9, no. 4 (October): 216-249.

Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics* 114, no. 2 (May): 533-575.

Angrist, Joshua D., Victor Lavy, Jetson Leder-Luis, and Adi Shany. 2019. "Maimonides' Rule Redux." *American Economic Review: Insights* 1, no. 3 (December): 309-324.

Argaw, Bethlehem A., and Patrick A. Puhani. 2018. "Does class size matter for school tracking outcomes after elementary school? Quasi-experimental evidence using administrative panel data from Germany." *Economics of Education Review* 65(August): 48-57.

Asadullah, M. Niaz. 2005. "The effect of class size on student achievement: evidence from Bangladesh." *Applied Economics Letters* 12, no. 4 (March): 217-221.

Atkinson, Adele, Simon Burgess, Bronwyn Croxson, Paul Gregg, Carol Propper, Helen Slater, and Deborah Wilson. 2009. "Evaluating the impact of performance-related pay for teachers in England." *Labour Economics* 16, no. 3: 251-261.

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

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

Barrera-Osorio, Felipe, Jacobus Cilliers, Marie-Hélène Cloutier, and Deon Filmer. 2022. "Heterogenous teacher effects of two incentive schemes: Evidence from a low-income country." *Journal of Development Economics* 156(May): 102820.

Barrera-Osorio, Felipe, and Dhushyanth Raju. 2017. "Teacher performance pay: Experimental evidence from Pakistan." *Journal of Public Economics* 148: 75-91.

Behrman, Jere R., Susan W. Parker, Petra E. Todd, and Kenneth I. Wolpin. 2015. "Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools." *Journal of Political Economy* 123, no. 2 (April): 325-364.

Bellés-Obrero, Cristina, and María Lombardi. 2022. "Teacher Performance Pay and Student Learning: Evidence from a Nationwide Program in Peru." *Economic Development and Cultural Change* 70, no. 4: 1631-1669.

Blimpo, Moussa P., David Evans, and Nathalie Lahire. 2015. "Parental Human Capital and Effective School Management: Evidence from The Gambia." Policy Research Working Paper 7238. Washington, DC: World Bank (April).

Bonesrønning, Hans. 2003. "Class Size Effects on Student Achievement in Norway: Patterns and Explanations." *Southern Economic Journal* 69, no. 4 (2003): 952-965.

Bosworth, Ryan. 2014. "Class size, class composition, and the distribution of student achievement." *Education Economics* 22, no. 2 (March): 141-165.

Brown, Christina, and Tahir Andrabi. 2021. "Inducing Positive Sorting through Performance Pay: Experimental Evidence from Pakistani Schools." University of California, Berkeley (January).
Browning, Martin, and Eskil Heinesen. 2007. "Class Size, Teacher Hours and Educational Attainment." *Scandinavian Journal of Economics* 109, no. 2: 415-438.
Brunner, Eric, Joshua Hyman, and Andrew Ju. 2020. "School Finance Reforms, Teachers' Unions, and the Allocation of School Resources." *The Review of Economics and Statistics* 102, no. 3: 473-489.
Buerger, Christian, Seung Hyeong Lee, and John D. Singleton. 2021. "Test-Based Accountability and the Eectiveness of School Finance Reforms." *AEA Papers and Proceedings* 111(May): 455-459.
Candelaria, Christopher A., and Kenneth A. Shores. 2019. "Court-Ordered Finance Reforms in the Adequacy Era: Heterogeneous Causal Effects and Sensitivity." *Education Finance and Policy* 14, no. 1 (Winter): 31-60.
Carlson, Deven, and Stéphane Lavertu. 2018. "School Improvement Grants in Ohio: Effects on Student Achievement and School Administration." *Educational Evaluation and Policy Analysis* 40, no. 3: 287-315.
Cascio, Elizabeth U., Nora Gordon, and Sarah Reber. 2013. "Local Responses to Federal Grants: Evidence from the Introduction of Title I in the South." *American Economic Journal: Economic Policy* 5, no. 3: 126-159.
Cho, Hyunkuk, Paul Glewwe, and Melissa Whitler. 2012. "Do reductions in class size raise students' test scores? Evidence from population variation in Minnesota's elementary schools." *Economics of Education Review* 31, no. 3: 77-95.
Clark, Melissa. 2003. "Education Reform, Redistribution, and Student Achievement: Evidence From the Kentucky Education Reform Act." (mimeo) Mathematica Policy Research (October).
Conlin, Michael, and Paul N. Thompson. 2017. "Impacts of new school facility construction: An analysis of a state-financed capital subsidy program in Ohio." *Economics of Education Review* 59(August): 13-28.
de Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers. 2018. "Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia." *Quarterly Journal of Economics* 133, no. 2: 993-1039.
Dee, Thomas S., and Benjamin J. Keys. 2004. "Does merit pay reward good teachers? Evidence from a randomized experiment." *Journal of Policy Analysis and Management* 23, no. 3 (Summer): 471-488.
Dee, Thomas S., and Martin R. West. 2011. "The non-cognitive returns to class size." *Educational Evaluation and Policy Analysis* 33, no. 1 (March): 23-46.
Denny, Kevin, and Veruska Oppedisano. 2013. "The surprising effect of larger class sizes: Evidence using two identification strategies." *Labour Economics* 23: 57-65.
Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives work: Getting teachers to come to school." *American Economic Review* 102, no. 4: 1241-1278.
Eren, Ozkan. 2019. "Teacher Incentives and Student Achievement: Evidence from an Advancement Program." *Journal of Policy Analysis and Management* 38, no. 4: 867-890.
Falch, Torberg, Astrid Marie Jorde Sandsør, and Bjarne Strøm. 2017. "Do Smaller Classes Always Improve Students' Long-run Outcomes?" *Oxford Bulletin of Economics and Statistics* 79, no. 5: 654-688.
Fredriksson, Peter, Björn Öckert, and Hessel Oosterbeek. 2013. "Long-Term Effects of Class Size." *Quarterly Journal of Economics* 128, no. 1: 249-285.
Fryer, Jr., Roland G., Steven D. Levitt, John List, and Sally Sadoff. 2022. "Enhancing the Efficacy of Teacher Incentives through Framing: A Field Experiment." *American Economic Journal: Economic Policy* 14, no. 4 (November): 269-99.
Fryer, Roland G. 2013. "Teacher incentives and student achievement: Evidence from New York City public schools." *Journal of Labor Economics* 31, no. 2: 373-427.

Gary-Bobo, Robert J., and Mohamed-Badrane Mahjoub. 2013. "Estimation of Class-Size Effects, Using "Maimonides' Rule" and Other Instruments: the Case of French Junior High Schools." *Annals of Economics and Statistics*, no. 111/112: 193-225.

Gigliotti, Philip, and Lucy C. Sorensen. 2018. "Educational resources and student achievement: Evidence from the Save Harmless provision in New York State." *Economics of Education Review* 66: 167-182.

Gilligan, Daniel O., Naureen Karachiwalla, Ibrahim Kasirye, Adrienne M. Lucas, and Derek Neal. 2022. "Educator Incentives and Educational Triage in Rural Primary Schools." *Journal of Human Resources* 57, no. 1: 79-111.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher incentives." *American Economic Journal: Applied Economics* 2, no. 3 (July): 205-227.

Goncalves, Felipe. 2015. "The Effects of School Construction on Student and District Outcomes: Evidence from a State-Funded Program in Ohio." (mimeo) Social Science Research Network (November 5).

Goodman, Sarena F., and Lesley J. Turner. 2013. "The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program." *Journal of Labor Economics* 31, no. 2 (April): 409-420.

Guryan, Jonathan. 2001. "Does Money Matter? Regression-Discontinuity Estimates from Education Finance Reform in Massachusetts." *National Bureau of Economic Research Working Paper Series* No. 8269.

Hægeland, Torbjørn, Oddbjørn Raaum, and Kjell G. Salvanes. 2012. "Pennies from heaven? Using exogenous tax variation to identify effects of school resources on pupil achievement." *Economics of Education Review* 31, no. 5: 601-614.

Hong, Kai. 2017. "School Bond Referendum, Capital Expenditure, and Student Achievement." *The B.E. Journal of Economic Analysis & Policy* 17, no. 4.

Hong, Kai, and Ron Zimmer. 2016. "Does Investing in School Capital Infrastructure Improve Student Achievement?" *Economics of Education Review* 53(August): 143-158.

Hoxby, Caroline M. 2000. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *Quarterly Journal of Economics* 115, no. 4 (November): 1239-1285.

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

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

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

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

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

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

Krueger, Alan B. 1999. "Experimental estimates of education production functions." *Quarterly Journal of Economics* 114, no. 2 (May): 497-532.

Krueger, Alan B., and Diane M. Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR." *The Economic Journal* 111, no. 468 (January): 1-28.

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

Lafortune, Julien, and David Schönholzer. 2022. "The Impact of School Facility Investments on Students and Homeowners: Evidence from Los Angeles." *American Economic Journal: Applied Economics* 14, no. 3: 254-89.

Lavy, Victor. 2002. "Evaluating the effect of teachers' group performance incentives on pupil achievement." *Journal of Political Economy* 110, no. 6 (December): 1286-1317.

Lavy, Victor. 2009. "Performance pay and teachers' effort, productivity, and grading ethics." *American Economic Review* 99, no. 5 (December): 1979-2011.

Leaver, Clare, Owen Ozier, Pieter Serneels, and Andrew Zeitlin. 2021. "Recruitment, Effort, and Retention Effects of Performance Contracts for Civil Servants: Experimental Evidence from Rwandan Primary Schools." *American Economic Review* 111, no. 7: 2213-46.

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

Leuven, Edwin, Mikael Lindahl, Hessel Oosterbeek, and Dinand Webbink. 2007. "The effect of extra funding for disadvantaged pupils on achievement." *Review of Economics and Statistics* 89, no. 4 (November): 721-736.

Leuven, Edwin, and Sturla A. Løkken. 2020. "Long-Term Impacts of Class Size in Compulsory School." *Journal of Human Resources* 55, no. 1 (January 1, 2020): 309-348.

Leuven, Edwin, Hessel Oosterbeek, and Marte Rønning. 2008. "Quasi-experimental Estimates of the Effect of Class Size on Achievement in Norway." *The Scandinavian Journal of Economics* 110, no. 4: 663-693.

Loyalka, Prashant, Sean Sylvia, Changfang Liu, James Chu, and Yaojiang Shi. 2019. "Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement." *Journal of Labor Economics* 37, no. 3.

Marsh, Julie A., Matthew G. Springer, Daniel F. McCaffrey, Kun Yuan, Scott Epstein, Julia Koppich, Nidhi Kalra, Catherine DiMartino, and Art (Xiao) Peng. 2011. *A Big Apple for Educators: New York City's Experiment with Schoolwide Performance Bonuses, Final Evaluation Report*. Santa Monica, CA: RAND Corporation.

Martorell, Paco, Kevin Stange, and Isaac McFarlin. 2016. "Investing in schools: capital spending, facility conditions, and student achievement." *Journal of Public Economics* 140(August): 13-29.

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2019. "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania." *Quarterly Journal of Economics* 134, no. 3: 1627-1673.

Mbiti, Isaac, Mauricio Romero, and Youdi Schipper. 2019. "Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania." *NBER Working Paper Series* No. 25903(May).

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

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher performance pay: Experimental evidence from India." *Journal of Political Economy* 119, no. 1 (February): 39-77.

Neilson, Christopher A., and Seth D. Zimmerman. 2014. "The effect of school construction on test scores, school enrollment, and home prices." *Journal of Public Economics* 120(December): 18-31.

Papke, Leslie E. 2008. "The Effects of Changes in Michigan's School Finance System." *Public Finance Review* 36, no. 4 (July): 456-474.

Pradhan, Menno, Daniel Suryadarma, Amanda Beatty, Maisy Wong, Arya Gaduh, Armida Alisjahbana, and Rima Prama Artha. 2014. "Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia." *American Economic Journal: Applied Economics* 6, no. 2: 105-126.

Rauscher, Emily. 2020a. "Delayed Benefits: Effects of California School District Bond Elections on Achievement by Socioeconomic Status." *Sociology of Education* 93, no. 2 (April): 110-131.

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

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

Roy, Joydeep. 2011. "Impact of School Finance Reform on Resource Equalization and Academic Performance: Evidence from Michigan." *Education Finance and Policy* 6, no. 2 (Spring): 137-167.

Schlaffer, James, and Gregory Burge. 2020. "The asymmetric effects of school facilities on academic achievement: Evidence from Texas bond votes." *The Social Science Journal*(April): 1-19.

Sojourner, Aaron J., Elton Mykerezi, and Kristine L. West. 2014. "Teacher Pay Reform and Productivity: Panel Data Evidence from Adoptions of Q- Comp in Minnesota." *Journal of Human Resources* 49, no. 4: 945-981.

Speroni, Cecilia, Alison Wellington, Paul Burkander, Hanley Chiang, Mariesa Herrmann, and Kristin Hallgren. 2020. "Do Educator Performance Incentives Help Students? Evidence from the Teacher Incentive Fund National Evaluation." *Journal of Labor Economics* 38, no. 3: 843-872.

Springer, Matthew G., Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J.R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher. 2010. *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching*. Nashville, TN: National Center on Performance Incentives, Vanderbilt University.

Urquiola, Miguel. 2006. "Identifying Class Size Effects in Developing Countries: Evidence from Rural Bolivia." *Review of Economics and Statistics* 88, no. 1: 171-177.

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

Winters, Marcus, Jay Greene, Gary Ritter, and Ryan Marsh. 2008. The Effect of Performance Pay in Little Rock, Arkansas on Student Achievement. Research Brief: National Center on Performance Incentives.

Wößmann, Ludger, and Martin West. 2006. "Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS." *European Economic Review* 50, no. 3: 695-736.

Zhang, Anwen. 2014. *Better Buildings, Better Scores? The Short-Run Effect of a Large School Construction Programme*, Working Paper.

Table 1: Distribution of funding source makeup with representative states, 2019 (percent)

Funding source	Mean	Minimum	Maximum
Local	42.26	2.10 (Hawaii)	91.97 (DC)
State	50.07	26.57 (Illinois)	90.29 (Vermont)
Federal	8.63	4.12 (New Jersey)	15.44 (Alaska)

Source: NCES 2021 digest (https://nces.ed.gov/programs/digest/2021menu_tables.asp), table 235.20

Table 2: NAEP and spending trends

Exam	Start year	End year	Δ score (SDs)	Δ score (SDs) per 10% spend inc.
Long term reading				
Age 9	1971	2012	0.3134	0.0266
Age 13	1971	2012	0.2135	0.0181
Age 17	1971	2012	0.0373	0.0032
Long term math				
Age 9	1978	2012	0.7049	0.0985
Age 13	1978	2012	0.5354	0.0748
Age 17	1978	2012	0.1705	0.0238
Reading				
Grade 4	1992	2019	0.1050	0.0247
Grade 8	1992	2019	0.0867	0.0204
Math				
Grade 4	1990	2019	0.8639	0.2028
Grade 8	1990	2019	0.5399	0.1268

Sources: Nation's Report Card (https://www.nationsreportcard.gov/) for main NAEP data and Long Term Trend NAEP data; NCES 2021 digest (https://nces.ed.gov/programs/digest/2021menu_tables.asp), table 236.55 for expenditure data

Notes: Δ score (SDs) reports the change in test scores in each respective exam over the period from Start year to End year in terms of the individual standard deviation of the exam in Start year. The next column reports this value for each 10% increase in national per-pupil expenditure (from the base level in Start year).

Table 3: Pandemic effect on NAEP scores

Exam	Start year	Δ score (SDs), 2019	Δ score (SDs), 2022
Reading			
Grade 4	1992	0.1050	0.0213
Grade 8	1992	0.0867	0.0120
Math			
Grade 4	1990	0.8639	0.7202
Grade 8	1990	0.5399	0.3252

Sources: Nation's Report Card (https://www.nationsreportcard.gov/) for main NAEP data Notes: Δ score (SDs), Year X reports the change in test scores in each respective exam over the period from Start year to Year X in terms of the individual standard deviation of the exam in Start year.

Table 4: School finance court cases by type and latest ruling

Type	Decision		Total
	For Plaintiff	For Defendant	
Equity	19	27	46
Adequacy	14	24	38
Both	60	54	114
Total	93	105	198

Source: Hanushek and Wirtz (forthcoming)
Notes: Some current cases are under appeal, and the decision refers to the last decision as of September 2022. Seven cases are not included because they did not have a final decision owing to a settlement or legislative action that ended the case. In general, the plaintiffs have brought suit to change the funding formula while the defendants represent the state government acting to stop the suit and to retain the current funding system.

Table 5: School finance court cases and baseline state expenditures

Type	Below natl. avg. PPE		Above natl. avg. PPE		Total
	For Plaintiff	For Defendant	For Plaintiff	For Defendant	
Equity	9	10	7	15	41
Adequacy	8	17	5	7	37
Both	28	25	31	23	107
Total	45	52	43	45	185

Source: Hanushek and Wirtz (forthcoming)
Notes: In the five-year period before the court filing, the average state expenditure per pupil is compared to the national average spending. In general, the plaintiffs have brought suit to change the funding formula while the defendants represent the state government acting to stop the suit and to retain the current funding system. Due to expenditure data availability constraints, 13 recent cases are omitted from this table. PPE = per-pupil expenditure.

Table 6: Percentage distribution of estimated effects of key resources on student performance, based on 376 studies

Resources	N	Statistically significant (%)		Statistically insignificant (%)
		Positive	Negative	
Real classroom resources				
Teacher-pupil ratio	276	14	14	72
Teacher education	170	9	5	86
Teacher experience	206	29	5	66
Financial aggregates				
Teacher salary	118	20	7	73
Expenditure per pupil	163	27	7	66
Other				
Facilities	91	9	5	86
Administration	75	12	5	83

Source: Hanushek (2003)

Table 7: Percentage distribution of estimated effect of teacher-pupil ratio and expenditure per pupil by state sampling scheme and aggregation

Level of aggregation of resources	N	Statistically significant (%)		Statistically insignificant (%)
		Positive	Negative	
Panel A: Teacher-pupil ratio				
Total	276	14	14	72
Single state samples[a]	157	11	18	71
Multiple state samples[b]	119	18	8	74
Disaggregated within states[c]	109	14	8	78
State level aggregation[d]	10	60	0	40
Panel B: Expenditure per pupil				
Total	163	27	7	66
Single state samples[a]	89	20	11	69
Multiple state samples[b]	74	35	1	64
Disaggregated within states[c]	46	17	0	83
State level aggregation[d]	28	64	4	32

Source: Hanushek (2003). Notes: a. Estimates from samples drawn within single states; b. Estimates from samples drawn across multiple states; c. Resource measures at level of classroom, school, district, or county, allowing for variation within each state; d. Resource measures aggregated to state level with no variation within each state.

Table 8: Percentage distribution of estimated influences on student performance, based on value-added models of individual student performance

Resources	N	Statistically significant (%)		Statistically insignificant (%)
		Positive	Negative	
Panel A: All studies				
Teacher-pupil ratio	78	12	8	80
Teacher education	40	0	10	90
Teacher experience	61	36	2	62
Panel B: Studies within a single state				
Teacher-pupil ratio	23	4	13	83
Teacher education	33	0	9	91
Teacher experience	36	39	3	58

Table 9: School spending estimate counts by study characteristics

Methodology	Outcome			Publication status	
	Test scores	Pass rates	Attainment	Published	Unpublished
	All studies				
DD	5	0	4	9	0
FE	0	1	0	1	0
IV	7	1	7	8	7
RD	7	0	7	11	3
RCT	4	0	0	3	1
	US studies only				
DD	5	0	4	9	0
FE	0	1	0	1	0
IV	6	1	7	7	7
RD	5	0	7	9	3
RCT	0	0	0	0	0

Notes: this table presents the counts of estimates (some studies produce multiple estimates) of the effects of school spending by outcome, publication status, and methodology. DD = difference-in-differences, FE = fixed effects, IV = instrumental variables, RD = regression discontinuity, RCT = randomized controlled trial

Table 10: Estimated impact of 10% increase per-pupil school spending, standardized parameter

Study	Unpublished	Outcome	Impact	SE	CI	Context
U.S. studies						
Abott, Kogan, Lavertu, and Peskowitz (2020)	X	Test scores	0.163	0.093	RD	Effects of narrow passage of Ohio tax referenda on test scores in grades 3-8 and 10
Abott, Kogan, Lavertu, and Peskowitz (2020)	X	Graduation	0.042	0.052	RD	Ohio tax referenda
Baron (2022)		Test scores	0.351	0.129	RD	Effects of narrow passage of Wisconsin expenditure referenda on 10th grade math scores; 1996-2014
Baron (2022)		Dropout	0.281	0.246	RD	Wisconsin expenditure referenda
Baron (2022)		College	0.418	0.139	RD	Wisconsin expenditure referenda
Baron, Hyman, and Vasquez (2022)	X	Test scores	0.011	0.025	IV	Michigan school funding reform; effect of 4 years of sustained funding increase on math exams
Baron, Hyman, and Vasquez (2022)	X	Graduation	0.034	0.011	IV	Michigan school funding reform; effect of 4 years of sustained funding increase on high school graduation rates
Baron, Hyman, and Vasquez (2022)	X	College	0.043	0.019	IV	Michigan school funding reform; effect of 4 years of sustained funding increase on college attendance rates
Brunner, Hyman, and Ju (2020)		Test scores	0.071	0.02	DD	School finance reforms across 13 states, 1986-2009; NAEP scores
Buerger, Lee, and Singleton (2021)		Test scores	0.069	0.034	DD	SFRs in 48 states, 1990-2011; 4th and 8th grade NAEP scores
Candelaria and Shores (2019)		Graduation	0.026	0.007	IV	SFRs, 1990-2010
Carlson and Lavertu (2018)		Test scores	0.135	0.071	RD	Ohio school improvement grants, 2009-2015; grades 3-8

Table 10 continued from previous page

Study	Unpublished	Outcome	Impact	SE	Causal inference	Context
Cascio, Gordon, Reber (2013)		Dropout	0.183	0.052	DD	Title I in poor districts in southern states, 1964-69
Clark (2003)	X	Test scores	0.054	0.044	IV	Kentucky funding and governance reform, 2000-2003; ACT scores
Gigliotti and Sorenson (2018)		Test scores	0.097	0.022	IV	New York State funding formula, 2007-2015; 3rd-8th grade math and reading test scores
Guryan (2001)	X	Test scores	0.087	0.034	IV	Massachusets funding reform, 1990-1997; 4th and 8th grade test scores
Hyman (2017)		College	0.066	0.031	IV	Michigan school funding reform, 2000-2017
Jackson, Johnson, and Persico (2016)		Graduation	0.082	0.017	IV	Long-term effects of SFRs, 1967-2010
Jackson, Wigger, and Xiong (2021)		Test scores	0.051	0.015	IV	Great recession spending cuts; NAEP scores in 4th and 8th grade
Jackson, Wigger, and Xiong (2021)		College	0.034	0.011	IV	Great recession spending cuts
Johnson (2015)		Graduation	0.129	0.055	DD	Long-term national impacts of Title I on cohorts born 1950-1970
Kreisman and Steinberg (2019)		Test scores	0.069	0.021	RD	Texas school funding formula, 2003-2010; grade 3-11 test scores
Kreisman and Steinberg (2019)		Dropout	0.316	0.118	RD	Texas school funding formula
Kreisman and Steinberg (2019)		Graduation	0.018	0.01	RD	Texas school funding formula
Kreisman and Steinberg (2019)		College	0.169	0.041	RD	Texas school funding formula
Lafortune, Rothstein, and Schanzenbach (2018)		Test scores	0.019	0.086	DD	Post-1990 SFRs; NAEP Test scores1990-2011
Lee and Polachek (2008)		Dropout	0.85	0.405	RD	New York Budget referenda

Table 10 continued from previous page

Study	Unpublished	Outcome	Impact	SE	Causal inference	Context
Miller (2018)	X	Test scores	0.077	0.02	IV	Changes in property values interacted with school finance formulas in 21 states, 2009-2013; 4th and 8th grade test scores
Miller (2018)	X	Graduation	0.047	0.012	IV	Changes in property values interacted with school finance formulas in 21 states, 2009-2013
Papke (2008)		Pass rates	0.059	0.008	FE	Michigan school funding reform; effect of 4 years of sustained funding increase on 4th grade math exams
Rauscher (2020b)		Test scores	0.543	0.2	DD	Funding reductions in rural districts in Kansas, 2010-2018
Rauscher (2020b)		Test scores	-0.244	0.22	DD	Funding reductions in non-rural districts in Kansas, 2010-2018
Rothstein and Schanzenbach (2022)		Test scores	0.012	0.005	DD	Effects of 4-year exposure to post-1990 adequacy reforms across the U.S.
Rothstein and Schanzenbach (2022)		College	0.011	0.006	DD	Effects of 4-year exposure to post-1990 adequacy reforms across the U.S.
Roy (2011)		Pass rates	0.054	0.022	IV	Michigan school funding reform; effect of 4 years of sustained funding increase on 4thgrade reading and math exams, 1998-2001
Weinstein, Stiefel, Schwartz, and Chalico (2009)	X	Test scores	-0.083	0.057	RD	Title I in NYC, 1997-2003; Test scores for grades 3-8
Developed nation studies						
Hægeland, Raaum, and Salvanes (2012)		Test scores	0.103	0.031	IV	Variation in taxable natural endowments feeding into average expenditure over 10 years in Norway, 1992-2003
Leuven et al. (2007)		Test scores	-0.182	0.093	RD	Unconditional teacher salary subsidies for disadvantaged Dutch schools, 1999-2003

Table 10 continued from previous page

Study	Unpublished	Outcome	Impact	SE	Causal inference	Context
Leuven et al. (2007)		Test scores	-0.118	0.124	RD	Classroom technology subsidies for disadvantaged Dutch schools, 1999-2003
Developing nation studies						
Blimpo, Evans, and Lahire (2015)	X	Test scores	-0.13	0.104	RCT	RCT in The Gambia with block grants, 2007-2011
de Ree et al. (2018)		Test scores	0.001	0.005	RCT	Unconditional teacher salary increase as RCT in Indonesia, 2009-2012
Mbiti et al. (2019)		Test scores	0.001	0.006	RCT	Unconditional grant RCT in Tanzanian primary schools, 2013-2014; year 2
Pradhan et al. (2014)		Test scores	0.144	0.191	RCT	RCT in Indonesia with block grants, 2007-2008

Notes: The estimates presented here have been scaled by the authors as detailed in Section 5 and Appendix Table 2 for the sake of reporting estimates in comparable terms. For test score estimates, results represent the effect of a 10% increase in spending on the change in test scores (in individual standard deviation units). For pass rates and all attainment outcomes, results represent the percent change in the outcome variable for a 10% increase in spending. For example, an estimate of .05 for graduation indicates that a 10% increase in spending led to a 5% increase in graduation rates. SFR = school finance reform.

Table 11: Distribution of standardized school spending estimates

Outcome	Median	Min	Max	N	N pos.	N Significant
Panel A: All studies (N=43)						
Test scores	0.069	-0.244	0.543	23	18	10
Pass rates	0.056	0.054	0.059	2	2	2
Attainment	0.057	0.011	0.850	18	18	14
Panel B: US studies only (N=36)						
Test scores	0.070	-0.244	0.543	16	14	9
Pass rates	0.056	0.054	0.059	2	2	2
Attainment	0.057	0.011	0.850	18	18	14

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

Table 12: School spending estimates by causal inference technique (test score outcomes only)

Causal inference	Median	Min	Max	N	N pos.	N Significant
Panel A: All studies (N=23)						
IV	0.077	0.011	0.103	7	7	5
RD	0.069	-0.182	0.351	7	4	2
DD	0.069	-0.244	0.543	5	4	3
RCT	0.001	-0.130	0.144	4	3	0
Panel B: US studies only (N=16)						
IV	0.066	0.011	0.097	6	6	4
DD	0.069	-0.244	0.543	5	4	3
RD	0.135	-0.083	0.351	5	4	2
RCT	-	-	-	-	-	-

Notes: The estimates presented here have been scaled by the authors as detailed in Section 5 and Appendix Table 2 for the sake of reporting estimates in comparable terms. The results represent the effect of a 10% increase in spending on the change in test scores (in individual standard deviation units). IV = instrumental variables, DD = difference-in-differences, RD = regression discontinuity design. Estimates are significant if $p < 0.05$.

Table 13: School spending estimates in U.S. by study characteristics (test score estimates only)

	Median	Min	Max	N	N pos.	N Significant
Panel A: single or multiple-state sample						
Within	0.078	-0.244	0.543	10	8	5
Across	0.070	0.019	0.163	6	6	4
Panel B: level of spending variation						
District	0.071	-0.244	0.543	13	12	8
School	0.026	-0.083	0.135	2	1	0
State	0.051	0.051	0.051	1	1	1
Panel C: level of outcomes data						
District	0.074	-0.244	0.543	12	11	8
Student	0.095	0.054	0.135	2	2	0
School	-0.083	-0.083	-0.083	1	0	0
State	0.051	0.051	0.051	1	1	1
Panel D: variation related to court-induced spending changes						
No	0.073	-0.244	0.543	10	8	5
Yes	0.070	0.019	0.097	6	6	4

Notes: The estimates presented here have been scaled by the authors as detailed in Section 5 and Appendix Table 2 for the sake of reporting estimates in comparable terms. The results represent the effect of a 10% increase in spending on the change in test scores (in individual standard deviation units). Within-state studies only look at students, schools, or districts within one state. In Panel A, across-state studies include data from several states, even if the test score and spending data are recorded at the district level. Thus, some across-state studies explicitly compare states while others simply include data from several states for increased sample size. Panel B considers the level of variation in spending as measured by the original study authors. If a study exploits differences in spending across states, the spending level is marked as "State." Panel C splits studies by the granularity of data used by study authors to measure test scores. If study authors use district average test scores as the outcome variable, the outcomes level would be marked as "District." In Panel D, court-induced spending changes include any policies that are directly related to school finance court cases as identified by the original study authors. These include both equity- and adequacy-related court-ordered school finance reforms. Estimates are significant if $p < 0.05$.

Table 14: Capital spending estimates

Study	Unpublished	Outcome	Impact	SE	CI	Context
U.S. studies						
Baron (2022)		Effect of bond passage on 10th grade math scores; average across first 10 post-election years	0.0125	0.0301	RD	Wisconsin bond referenda averaging $8,200 pp, 1996-2015
Baron (2022)		Effect of bond passage on dropout rate; average across first 10 post-election years	0.0808	0.0808	RD	Wisconsin bond referenda averaging $8,200 pp, 1996-2015
Baron, Hyman, and Vasquez (2022)	X	Effect of bond passage in K on 4th and 7th grade math scores	0.0114	0.0437	RD	Michigan bond referenda
Baron, Hyman, and Vasquez (2022)	X	Effect of bond passage in K on high school graduation	0.0024	0.0003	RD	Michigan bond referenda
Cellini, Ferreira, and Rothstein (2010)		Effect of bond passage on test scores, 6 years post	0.0719	0.0336	RD	California bond referenda averaging $11,600 pp, 1987-2006; 3rd and 4th grade
Conlin and Thompson (2017)		Effect of a $1200 increase in pp capital expenditures 3 years prior on proficiency rates	-0.0005	0.0006	IV	Ohio grants for school facility upgrades, 1997-2011
Conlin and Thompson (2017)		Effect of a $1200 increase in the value of pp capital stock 4 years prior on proficiency rates	0.0004	0.0001	IV	Ohio grants for school facility upgrades, 1997-2011
Goncalves (2015)	X	Effect of exposure to construction on proficiency rates; 4 years post	-0.0267	0.0093	FE	Ohio grants for school facility upgrades, 1998-2014
Goncalves (2015)	X	Effect of exposure to completed construction; 6 years post	-0.0011	0.0131	FE	Ohio grants for school facility upgrades, 1998-2014
Hong (2017)		Effect of bond passage on 4th and 7th grade reading proficiency, 6 years post	-0.0009	0.0069	RD	Michigan bond referenda averaging $13,151 pp, 1996-2009

Table 14 continued from previous page

Study	Unpublished	Outcome	Impact	SE	CI	Context
Hong and Zimmer (2016)		Effect of bond passage on 4th and 7th grade reading proficiency, 6 years post	0.0289	0.0165	RD	Michigan bond referenda averaging $13,151 pp, 1996-2009
Lafortune and Schonholzer (2019)		Effect of an additional year of exposure to a newly constructed schools on test scores	0.029	0.0104	IV	School construction program in Los Angeles averaging $92,300 per seat, 1997-2008
Martorell, Stange, and McFarlin (2016)		Effect of bond passage on test scores in grades 3-8 and 10 , 6 years post	0.0075	0.0156	RD	Texas bond referenda averaging $14,200 pp, 1997-2010
Martorell, Stange, and McFarlan (2016)		Effect of bond passage on attendance rate in grades 3-8, 6 years post	-0.0001	0.0007	RD	Texas bond referenda averaging $14,200 pp, 1997-2010
Neilson and Zimmerman (2014)		Effect of new school construction on test scores in grades 3-8, 6 years post occupancy	0.092	0.0585	FE	School construction program in New Haven, CT, 2002-2010
Rauscher (2020b)		Effect of bond passage on high SES test scores, 6 years post	0.1498	0.2011	RD	California bond referenda with average close election measure concerning $9,700 in per-pupil revenue
Rauscher (2020b)		Effect of bond passage on low SES test scores, 6 years post	0.579	0.2878	RD	California bond referenda with average close election measure concerning $9,700 in per-pupil revenue
Schlaffer and Burge (2020)		Effect of bond passage on scores in grades 3-8, 6 years post	0.063	0.017	RD	Texas bond referenda, 2003-2014
Schlaffer and Burge (2020)		Effect of new school construction on test scores in grades 3-8, 4+ years post occupancy	0.0928	0.0079	DD	Texas school construction; 2003-2014
Developed nation studies						

Table 14 continued from previous page

Study	Unpublished	Outcome	Impact	SE	CI	Context
Zhang (2014)	X	Effect of new school construction on grade 11 GCSE scores, 3 years post	0.1304	0.0811	DD	England school construction program averaging $17k per pupil for new schools, 2003-2010

Notes: The estimates presented here have been scaled by the authors as detailed in Section 5 and Appendix Table 3 for the sake of reporting estimates in comparable terms. Test score estimates are scaled in terms of individual standard deviations and pass rate or attainment estimates are scaled in terms of % Δ in base levels. Because magnitude of spending changes differ greatly across studies, the Context column provides information on dollar values if available.

Table 15: Estimated effects of a 1-student reduction in class size, standardized parameter

Study	Outcome	Impact	SE	ID	Context
US Studies					
Bosworth (2014)	Test scores	0.0042	0.0003	FE	4th and 5th grade students in North Carolina public schools, 2001-2004
Cho, Glewwe, and Whitler (2005)	Test scores	0.0044	0.0016	Random enrollment variation	3rd and 5th grade students in Minnesota, 1997-2005
Dee and West (2011)	Test scores	0.0018	0.0021	Within-student variation in size across subjects	Nationally representative sample of American 8th grade students, 1988
Denny and Oppedisano (2013)	Test scores	-0.017	0.0215	Random enrollment variation	Student performance on the 2003 PISA in the US
Hoxby (2000)	Test scores	0.0017	0.0036	Random enrollment variation	Connecticut elementary schools, 1992-1998
Jepsen and Rivkin (2009)	Test scores	0.0058	0.0004	FE	California class size reduction program in primary school, 1990-2002
Krueger (1999)	Test scores	0.0159	0.0043	RCT: Project STAR	Standardized test score effects of on year of random assignment to a smaller class in Tennessee, grades K-3 and cohorts born 1985-1989
Krueger and Whitmore (2001)	Test scores	0.029	0.0095	RCT: Project STAR	ACT/SAT score effects of random assignment to smaller classes in K-3 in Tennesse, cohorts starting school 1985-1989
Developed nation studies					
Angrist and Lavy (1999)	Test scores	0.0139	0.0041	Max. class size rule	3rd-5th grade classes in Israel with a maximum size rule of 40, 1991-1992
Angrist, Battistin, and Vuri (2017)	Test scores	-0.001	0.0017	Max. class size rule	Italian 2nd and 5th grade classes, 2009-2012; corrected for test score manipulation
Angrist, Lavy, Leder-Luis, and Shany (2019)	Test scores	0.0005	0.0017	Max. class size rule	Israeli 5th graders, 2002-2011
Argaw and Puhani (2018)	Tracking	0.0056	0.0037	Max. class size rule	4th grade students in the German state of Hesse, 2007-2013

Table 15 continued from previous page

Study	Outcome	Impact	SE	ID	Context
Bønesronning (2003)	Test scores	0.0131	0.0088	Max. class size rule	Exploiting a maximum class size rule of 30 for Norwegian 8th-10th grade students
Browning and Heinesen (2007)	Years of education	0.0016	0.0012	Max. class size rule	Long-term effects of class size in 8th grade a maximum class size of 24, Danish cohorts born 1971-1978
Denny and Oppedisano (2013)	Test scores	-0.0655	0.0331	Random enrollment variation	Student performance on the 2003 PISA in the UK
Falch, Sandsor, and Strom (2017)	Years of education	-0.0001	0.0001	Max. class size rule	Long-term effects of average class size experienced by a cohort in 8th-10th grade under a maximum class size of 30, Norwegian cohorts born 1966-1984
Fredriksson, Ockert, and Oosterbeek (2013)	Test scores	0.0233	0.0101	Max. class size rule	Effects of average class size at ages 10-13 on Test scoresat age 16 in Sweden, 1977-1995
Gary-Bobo and Mahjoub (2013)	Grade promotion	0.0028	0.0051	Random enrollment variation	Effect of class size on year-end grade promotion in Fench junior high schools, 1989-1993
Leuven and Løkken (2018)	Years of education	0.0006	0.0007	Max. class size rule	Long-term effects of class size in Norway using a maximum class size of 28, cohorts born after 1978
Leuven, Oosterbeek, and Ronning (2008)	Test scores	-0.005	0.0044	Max. class size rule	Effects of average class size experieneced by a cohort over three years for Norwegian students in grades 7-9, 2001-2003
Leuven, Oosterbeek, and Ronning (2008)	Test scores	-0.0082	0.0144	Random enrollment variation	Effects of average class size experieneced by a cohort over three years for Norwegian students in grades 7-9, 2001-2003
Wößmann and West (2006)	Test scores	-0.0097	0.0119	Random enrollment variation	TIMSS scores in 7th and 8th grade: Belgium
Wößmann and West (2006)	Test scores	-0.003	0.0072	Random enrollment variation	TIMSS scores in 7th and 8th grade: Canada
Wößmann and West (2006)	Test scores	-0.0282	0.0238	Random enrollment variation	TIMSS scores in 7th and 8th grade: Czech Republic

Table 15 continued from previous page

Study	Outcome	Impact	SE	ID	Context
Wößmann and West (2006)	Test scores	0.0347	0.0174	Random enrollment variation	TIMSS scores in 7th and 8th grade: France
Wößmann and West (2006)	Test scores	0.0171	0.0111	Random enrollment variation	TIMSS scores in 7th and 8th grade: Greece
Wößmann and West (2006)	Test scores	0.0363	0.0119	Random enrollment variation	TIMSS scores in 7th and 8th grade: Iceland
Wößmann and West (2006)	Test scores	-0.0242	0.011	Random enrollment variation	TIMSS scores in 7th and 8th grade: Portugal
Wößmann and West (2006)	Test scores	0.0034	0.0191	Random enrollment variation	TIMSS scores in 7th and 8th grade: Romania
Wößmann and West (2006)	Test scores	-0.0049	0.0054	Random enrollment variation	TIMSS scores in 7th and 8th grade: Singapore
Wößmann and West (2006)	Test scores	-0.0141	0.0164	Random enrollment variation	TIMSS scores in 7th and 8th grade: Slovenia
Wößmann and West (2006)	Test scores	0.0042	0.0116	Random enrollment variation	TIMSS scores in 7th and 8th grade: Spain
Developing nation studies					
Asadullah (2005)	Pass rates	-0.1663	0.0489	Max. class size rule	Exploiting a max class size rule of 60 for Bangladeshi 10th graders, 1999
Urquiola (2006)	Test scores	0.0313	0.0174	Max. class size rule	Exploiting a maximum class size rule of 30 for Bolivian 3rd grade students

Notes: The estimates of effects of 1-student reductions in class size presented here have been scaled by the authors as detailed in Section 5 and Appendix Table 4 for the sake of reporting estimates in comparable terms. Test score estimates are scaled in terms of individual standard deviations and pass rate or attainment estimates are scaled in terms of % Δ in base levels. All studies are published in peer-reviewed journals.

Table 16: Distribution of class size estimates

Outcome	Median	Min	Max	N	N pos.	N Significant
Panel A: All studies (N=33)						
Test scores	0.003	-0.066	0.036	29	18	9
Pass rates	-0.166	-0.166	-0.166	1	0	0
Attainment	0.011	-0.000	0.006	4	3	0
Panel B: US studies only (N=8)						
Test scores	0.004	-0.017	0.029	8	7	5
Pass rates	-	-	-	-	-	-
Attainment	-	-	-	-	-	-

Notes: The estimates of effects of 1-student reductions in class size presented here have been scaled by the authors as detailed in Section 5 and Appendix Table 4 for the sake of reporting estimates in comparable terms. Test score estimates are scaled in terms of individual standard deviations and pass rate or attainment estimates are scaled in terms of % Δ in base levels. Estimates are significant if $p < 0.05$.

Table 17: Estimates of the effect of switching to performance-related pay, standardized parameter

Study	Unpub.	Outcome	Impact	SE	CI	Level	Style	Context	Salary %
U.S. studies									
Dee and Keys (2004)		Test scores	0.104	0.044	IV	Individual	Piece-rate	Multi-level career ladder with monetary incentives for Tennessee teachers of K-3 students as part of the 1985-1989 STAR program	NA
Eren (2019)		Test scores	0.106	0.076	DD	Mix	Mix	Hybrid P4P program in Louisiana schools; analysis focuses on elementary and middle schools in 2003-2005	5
Fryer (2013)		Test scores	-0.02	0.01	RCT	Individual	Piece-rate	RCT with high-poverty public schools in NYC, 2007-2010	NA
Fryer, Levitt, List, and Sadoff (2022)		Test scores	0.124	0.056	RCT	Individual	Tournament	Loss aversion RCT with initial bonuses in elementary and middle schools in suburban district near Chicago, 2010-2012	8
Fryer, Levitt, List, and Sadoff (2012)		Test scores	0.051	0.062	RCT	Individual	Tournament	RCT with pay-for-percentile year-end bonuses in elementary and middle schools in suburban district near Chicago, 2010-2012	8
Goodman and Turner (2013)		Test scores	-0.016	0.012	IV	Group	Piece-rate	Treatment-on-treated effects from RCT providing P4P bonuses to elementary and middle schools in NYC, 2007-9	5
Marsh et al. (2011)	X	Test scores	-0.02	0.02	RCT	Group	Piece-rate	Merit pay RCT in high-need NYC elementary, middle, and high schools, 2007-2010, math scores	5

Table 17 continued from previous page

Study	Unpub.	Outcome	Impact	SE	CI	Level	Style	Context	Salary %
Sojourner, Mkerezi, and West (2014)		Test scores	0.011	0.011	DD	Individual	Piece-rate	P4P reform in Minnesota; outcomes for grades 3-8, 2003-2009	NA
Speroni et al. (2020)		Test scores	0.048	0.016	RCT	Individual	Mix	Random assignment to P4P through the Teacher Incentive Fund grant program across several US states, 2011-15	5
Springer et al. (2010)	X	Test scores	0.03	0.02	RCT	Individual	Piece-rate	Nashville test score gains-based P4P randomized program in 2006-9 covering elementary and middle schools	NA
Winters et al. (2008)	X	Test scores	0.158	0.053	DD	Individual	Piece-rate	District wide P4P program in Little Rock, Arkansas elementary schools, 2004-2007	NA
Developed nation studies									
Atkinson et al. (2009)		Test scores	0.53	0.22	DD	Individual	Piece-rate	Widespread salary progression P4P in UK schools, 1997-2002	NA
Behrman, Parker, Todd, and Wolpin (2015)		Test scores	-0.001	0.034	RCT	Individual	Piece-rate	Gains-based incentives for math teachers in grades 10-12 in Mexican federal high schools; years 2008-2011	3
Lavy (2002)		Test scores	0.083	0.04	RD	Group	Tournament	School-level merit pay enacted for Israeli high schools that were the only one of their kind in their community in 1996 and 1997	2
Lavy (2009)		Pass rates	0.093	0.047	DD	Individual	Tournament	P4P in low-performing Israeli high schools, 1999-2001	13

Table 17 continued from previous page

Study	Unpub.	Outcome	Impact	SE	CI	Level	Style	Context	Salary %
Loyalka et al. (2019)		Test scores	0.074	0.044	RCT	Individual	Tournament	RCT in Chinese primary schools using bonuses to reward math instructors for test score performance, 2012-2014	12
Obrero and Lombardi (2021)		Grades	-0.001	0.006	DD	Group	Tournament	Large-scale P4P bonus program in Peruvian secondary schools based on test score levels, 2013-2015	10
Developing nation studies									
Andrabi and Brown (2021)	X	Test scores	0.091	0.058	RCT	Individual	Tournament	RCT in Pakistani schools, objective score-based incentives, 2017-19	6
Brown and Andrabi (2022)	X	Test scores	0.088	0.04	RCT	Individual	Tournament	Contract choice and RCT for assignment to P4P in Pakistani private schools	6
Barrera-Osorio and Raju (2017)		Test scores	0.008	0.06	RCT	Group	Piece-rate	RCT in Pakistani primary schools with lowest base test scores, 2010-2013	8
Barrera-Osorio et al. (2022)		Test scores	0.239	0.084	RCT	Individual	Piece-rate	RCT providing in-kind bonuses to Guinean primary school teachers for test score gains, 2012-14	27
Duflo, Hanna, and Ryan (2012)		Test scores	0.17	0.06	RCT	Individual	Piece-rate	RCT with contracts incentivizing teacher attendance in India	31.5
Gilligan et al. (2019)		Test scores	0.018	0.03	RCT	Individual	Tournament	RCT for Ugandan 6th grade math teachers, 2016-2018	8
Gilligan et al. (2019)		Attendance	0.075	0.032	RCT	Individual	Tournament	RCT for Ugandan 6th grade math teachers, 2016-2018	8

Table 17 continued from previous page

Study	Unpub.	Outcome	Impact	SE	CI	Level	Style	Context	Salary %
Glewwe, Ilias, and Kremer (2010)		Test scores	0.048	0.061	RCT	Group	Tournament	RCT in Kenyan primary schools with gift bonuses rewarding school-wide absolute and relative test score performance of 4th-8th grade students	NA
Leaver et al. (2020)	X	Test scores	0.06	0.17	RCT	Individual	Tournament	RCT in Rwandan primary schools randomizing pay structure at both recruitment and teaching stages, 2016	15
Mbiti et al. (2019)		Test scores	0.21	0.07	RCT	Individual	Piece-rate	RCT in Tanzanian primary schools, 2013-2014	NA
Mbiti, Romero, and Scipper (2019)	X	Test scores	0.17	0.064	RCT	Individual	Piece-rate	2014-2016 experiment in Tanzania comparing pay-for-percentile and levels incentives in elementary schools	4
Mbiti, Romero, and Scipper (2019)	X	Test scores	0.059	0.054	RCT	Individual	Tournament	2014-2016 experiment in Tanzania comparing pay-for-percentile and levels incentives in elementary schools	4
Muralidharan and Sundararaman (2011)		Test scores	0.156	0.05	RCT	Individual	Piece-rate	RCT with test score gains-based bonuses in Indian primary schools, 2005-2007	NA
Muralidharan and Sundararaman (2011)		Test scores	0.141	0.05	RCT	Group	Piece-rate	RCT with test score gains-based bonuses in Indian primary schools, 2005-2007	NA

Notes: The estimates of effects of switching to performance-related pay presented here have been scaled by the authors as detailed in Section 5 and Appendix Table 5 for the sake of reporting estimates in comparable terms. Test score estimates are scaled in terms of individual standard deviations and pass rate or attainment estimates are scaled in terms of % Δ in base levels. For those studies with information regarding the size of bonuses or incentive pay, the last column presents the value of the additional average yearly incentive pay as a percentage of the average yearly base teacher salary. P4P = pay for performance.

Table 18: Distribution of performance pay estimates

Outcome	Median	Min	Max	N	N pos.	N Significant
Panel A: All studies (N=31)						
Test scores	0.074	-0.020	0.530	29	24	13
Pass rates	0.093	0.093	0.093	1	1	1
Attainment	0.075	0.075	0.075	1	1	1
Panel B: US studies only (N=11)						
Test scores	0.048	-0.020	0.158	11	8	4
Pass rates	-	-	-	-	-	-
Attainment	-	-	-	-	-	-

Notes: The estimates of effects of switching to performance-related pay presented here have been scaled by the authors as detailed in Section 5 and Appendix Table 5 for the sake of reporting estimates in comparable terms. Test score estimates are scaled in terms of individual standard deviations and pass rate or attainment estimates are scaled in terms of % Δ in base levels. Estimates are significant if $p < 0.05$.

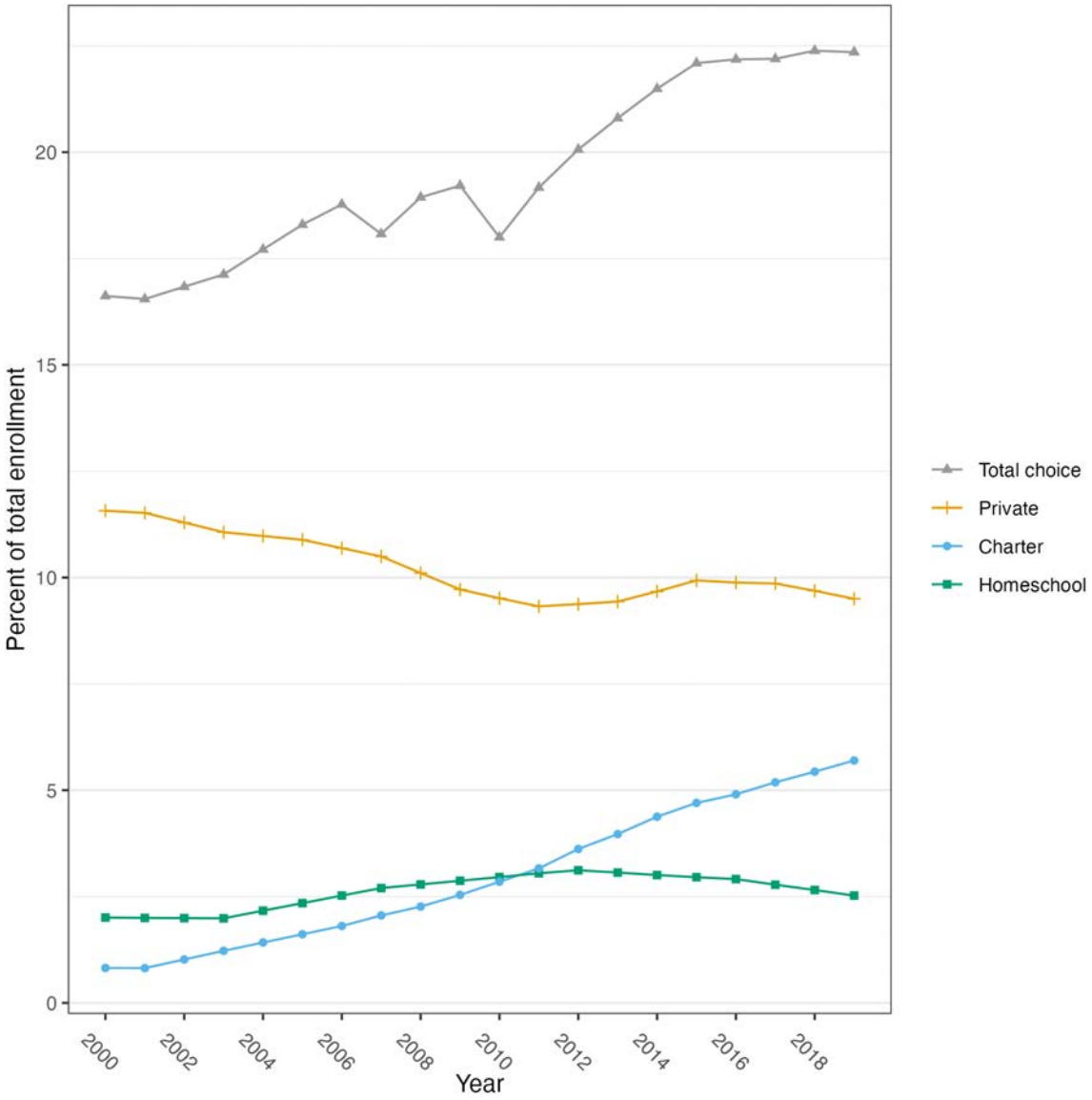

Figure 1: Enrollment patterns of U.S. children, 2000-2019

Source: Author calculations using NCES 2021 digest (https://nces.ed.gov/programs/digest/2021menu_tables.asp), tables 205.10, 206.10, 203.20, and 216.20 as well as NCES digest 2014 table 216.20.

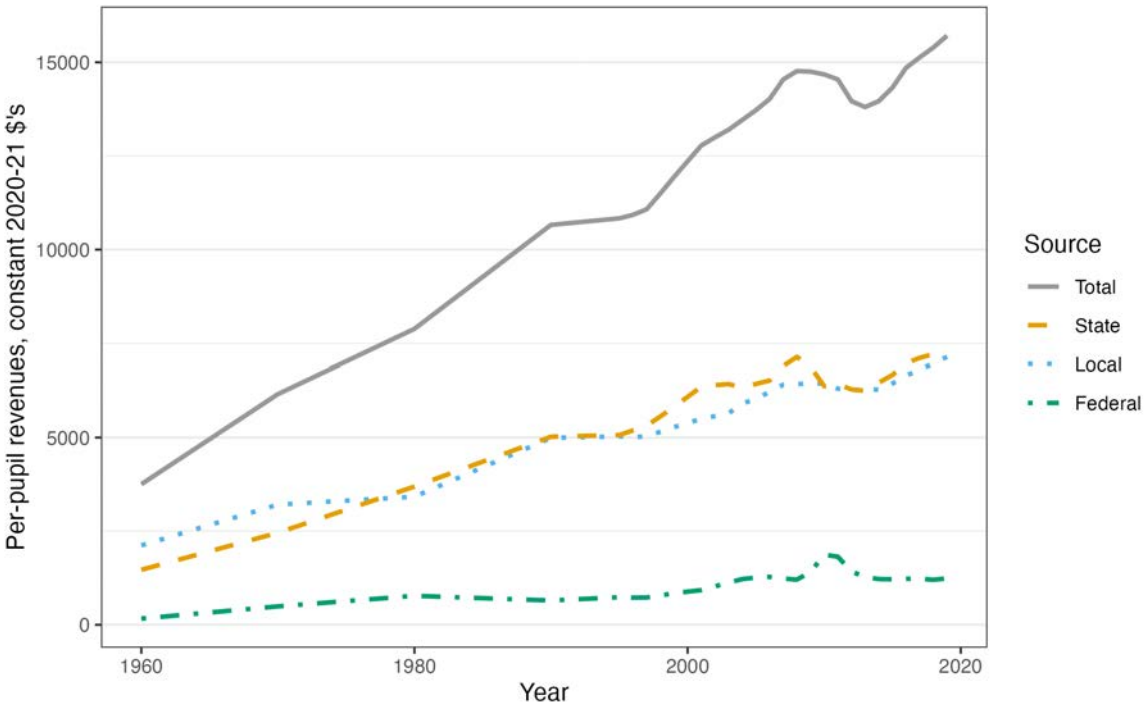

Figure 2: Per-pupil revenue of U.S. public schools, by source, 1960-2019

Source: NCES 2021 digest (https://nces.ed.gov/programs/digest/2021menu_tables.asp), table 235.10

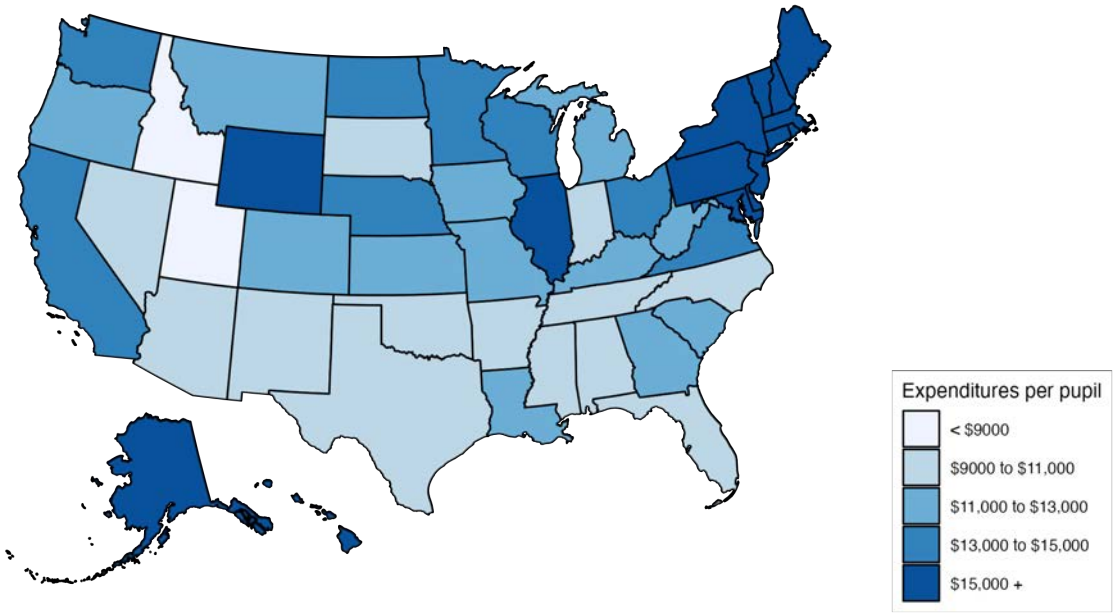

Figure 3: Per-pupil expenditure by state, 2019

Source: NCES 2021 digest (https://nces.ed.gov/programs/digest/2021menu_tables.asp), table 236.65

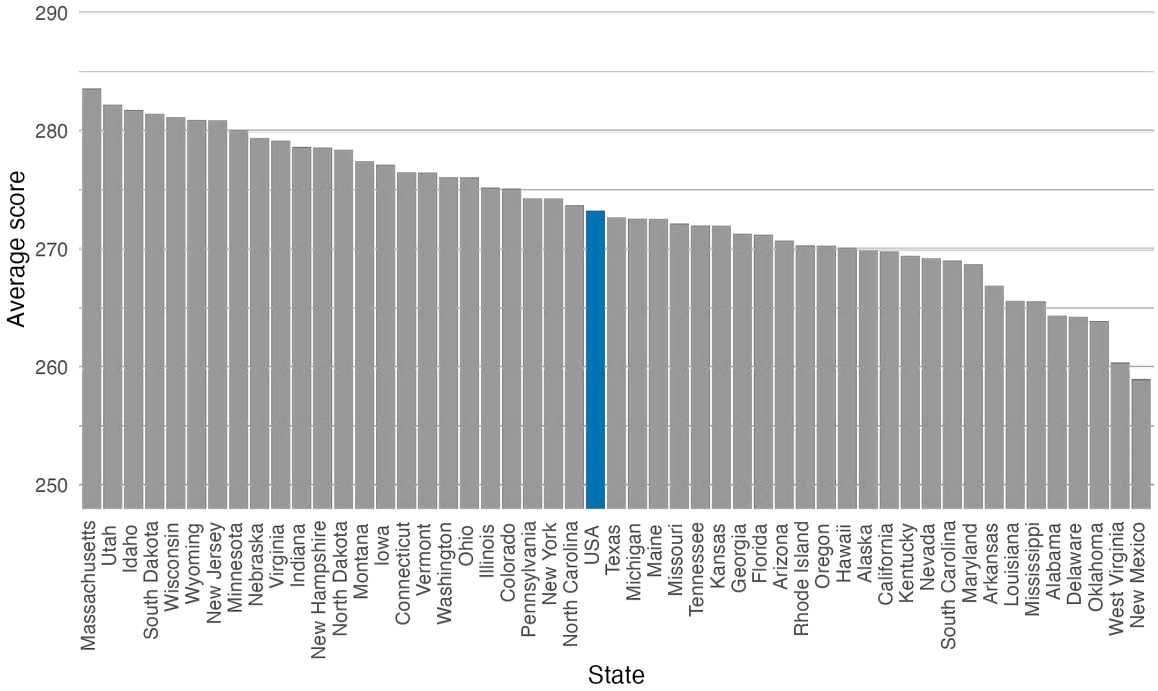

Figure 4: Grade 8 National Assessment of Educational Progress (NAEP) scores by state, 2022

Source: Nation's Report Card (https://www.nationsreportcard.gov/)
Notes: National standard deviation of individual scores = 39

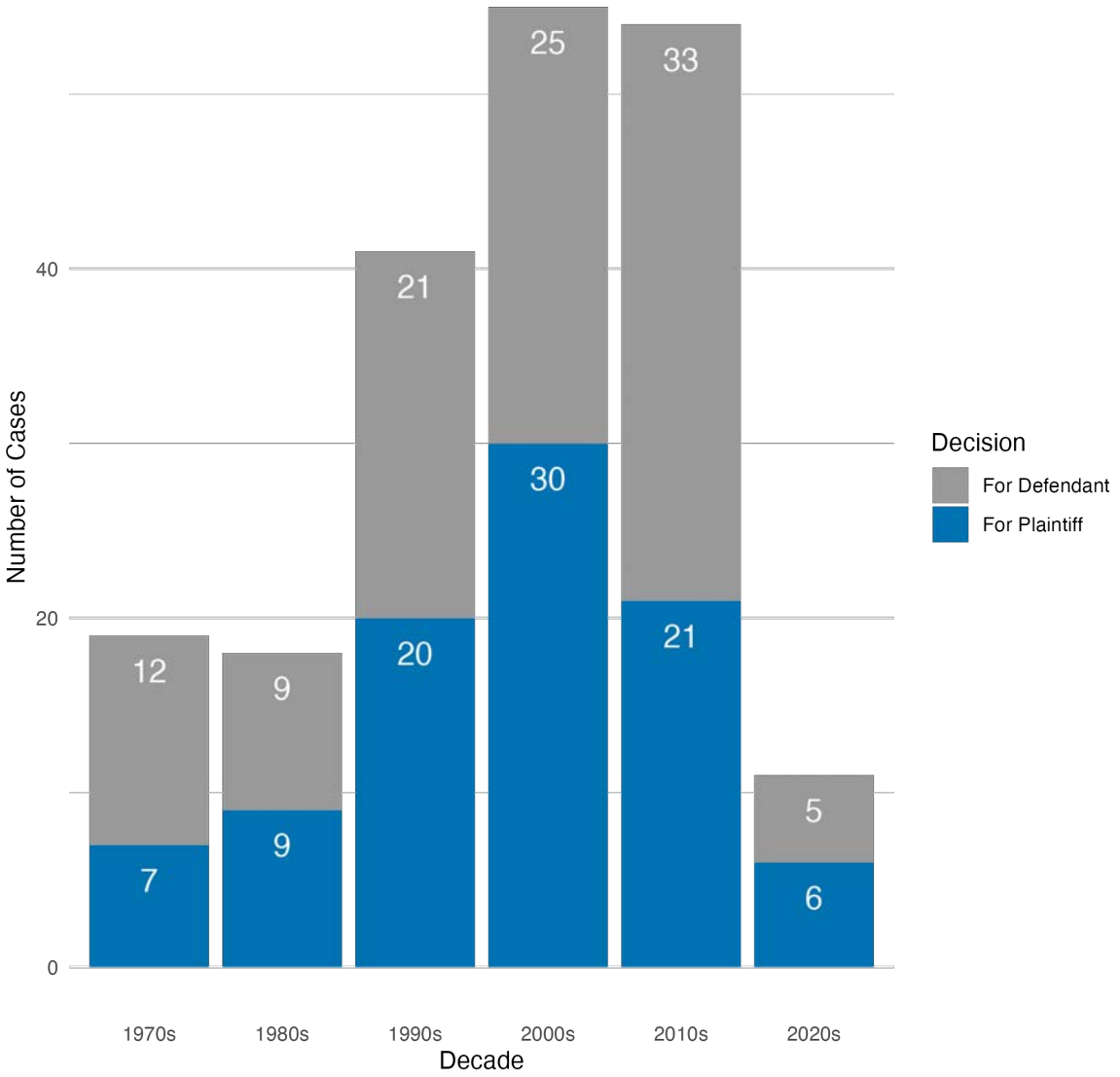

Figure 5: Type of school finance court case by decade

Source: Hanushek and Wirtz (forthcoming)
Notes: Some current cases are under appeal, and the decision refers to the last decision as of September 2022. Seven cases are not included because they did not have a final decision owing to a settlement or legislative action that ended the case. In general, the plaintiffs have brought suit to change the funding formula while the defendants represent the state government acting to stop the suit and to retain the current funding system.

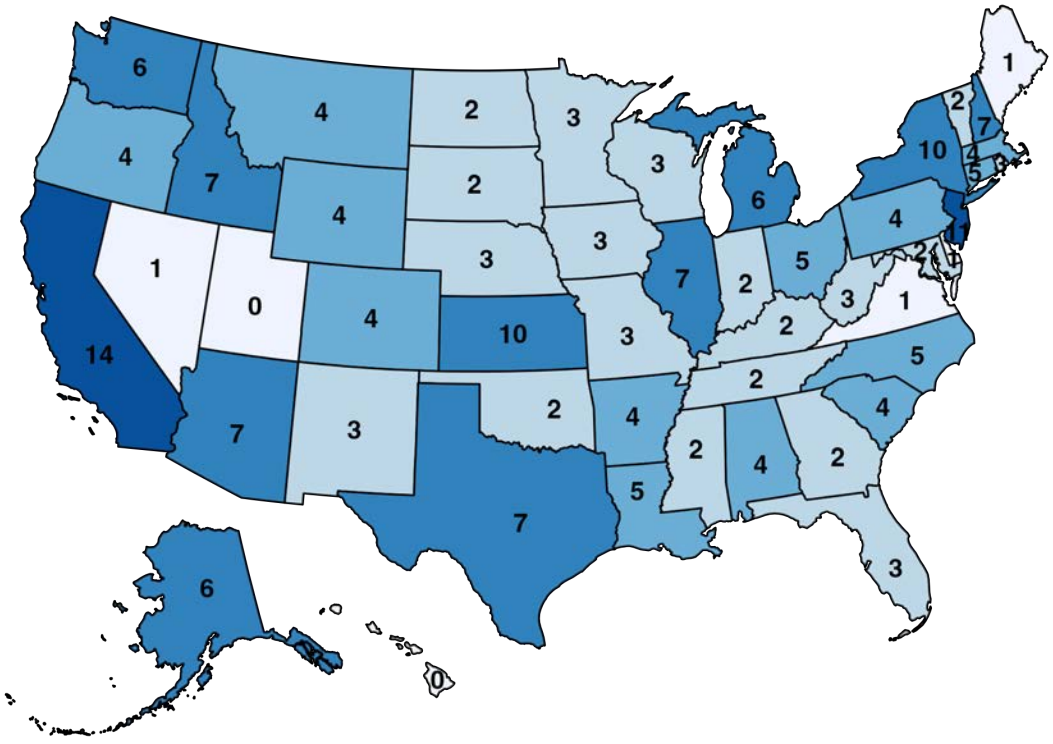

Figure 6: Number of cases per state

Source: Hanushek and Wirtz (forthcoming)

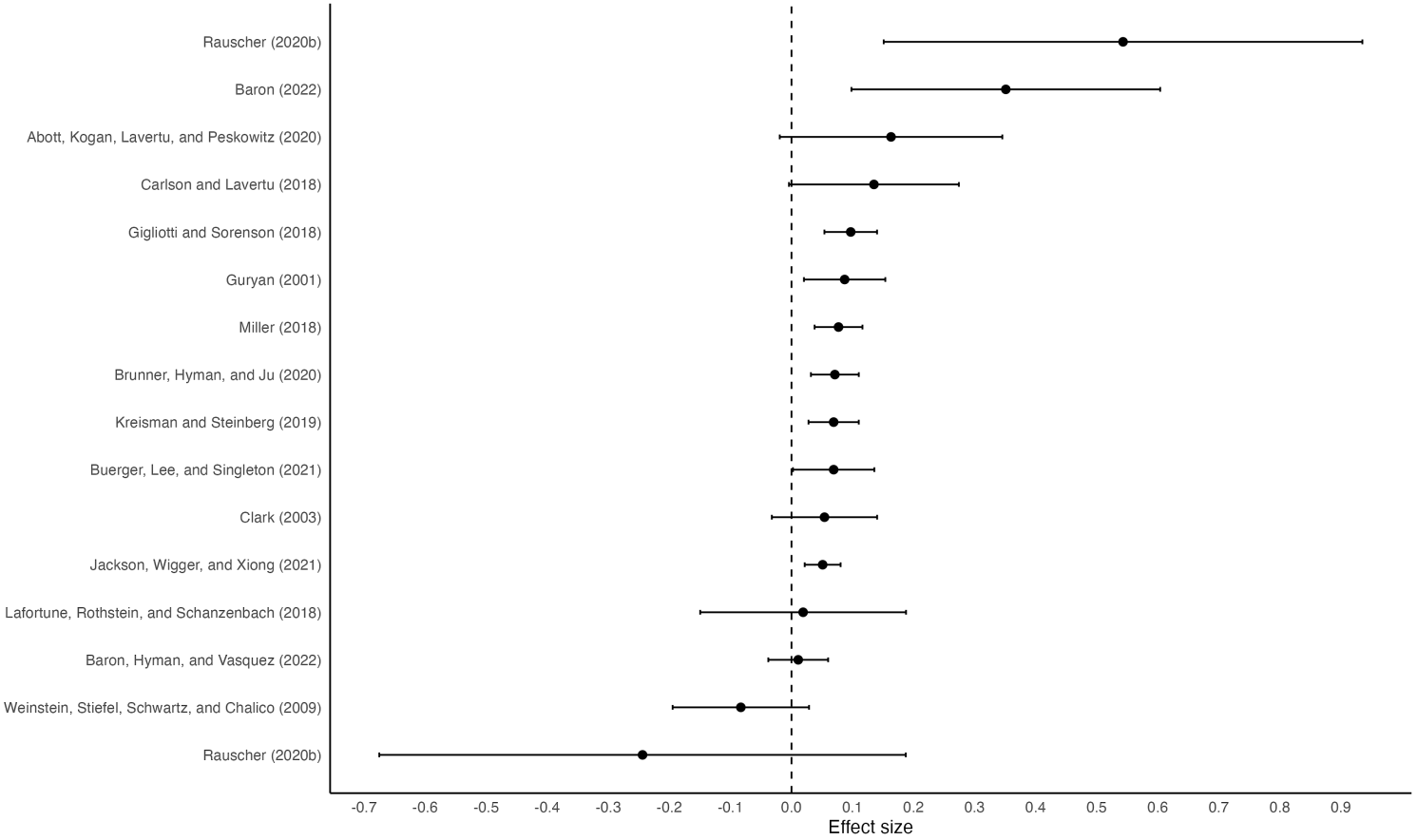

Figure 7: Effects of school spending on test scores, US

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

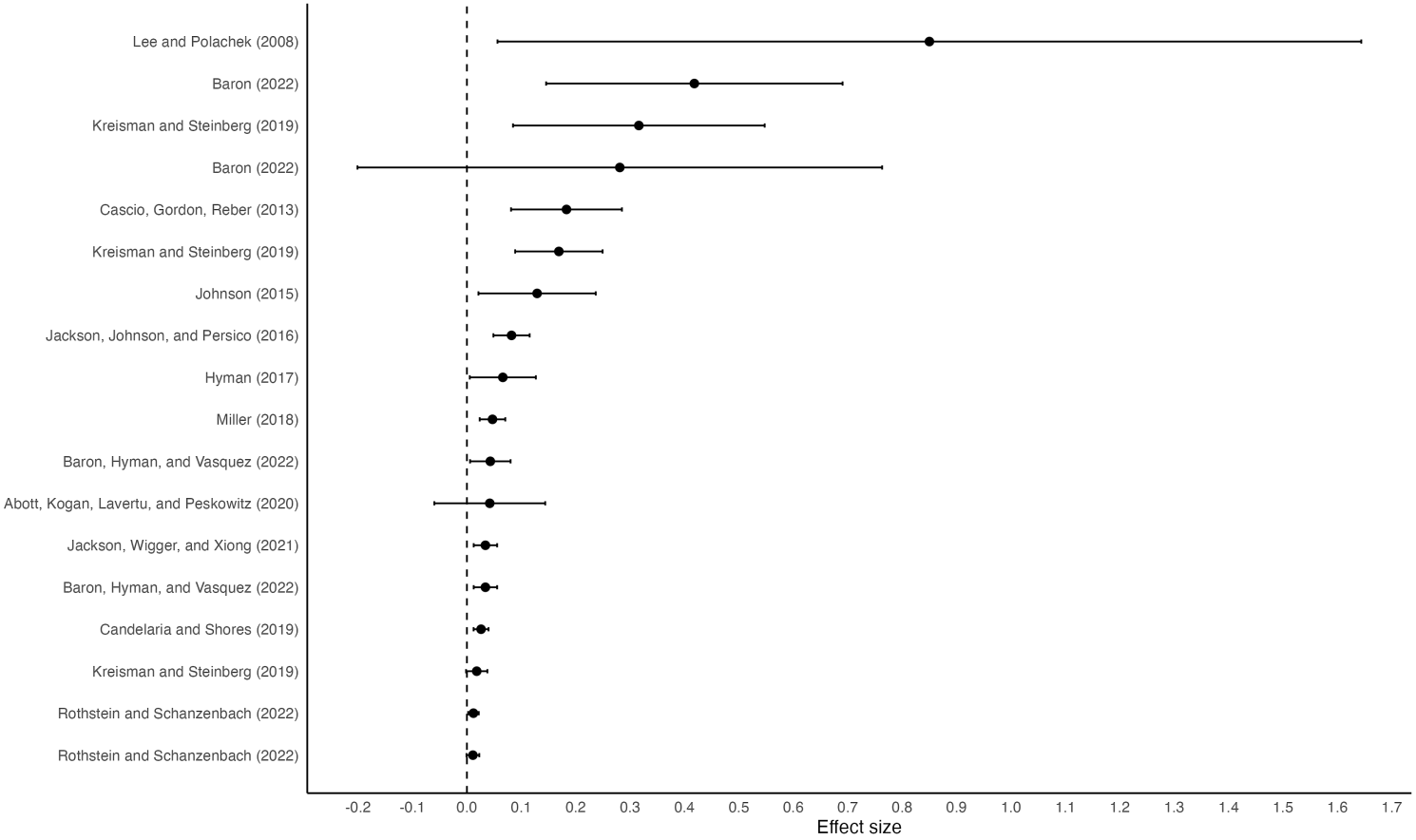

Figure 8: Effects of school spending on attainment, US

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

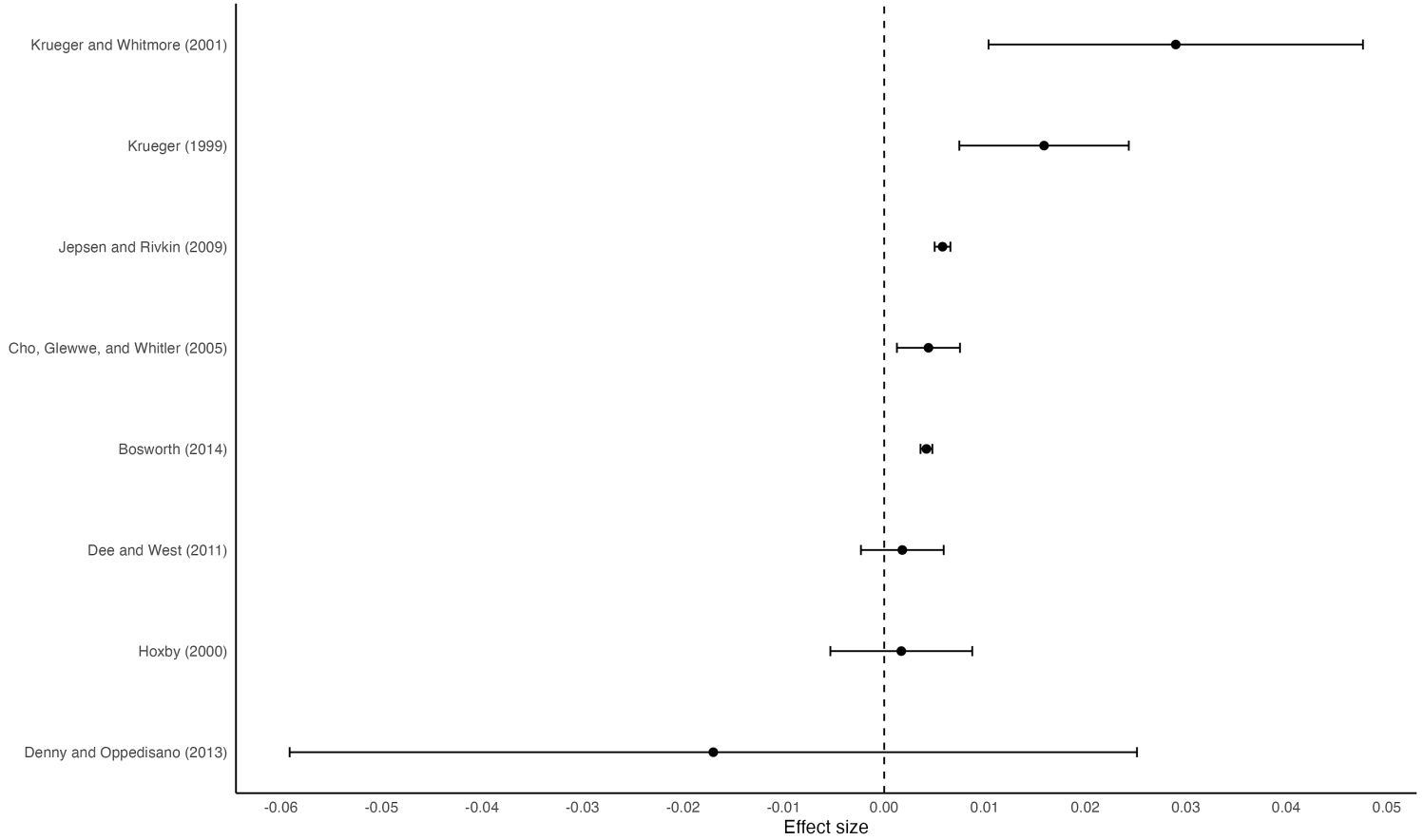

Figure 9: Effects of class size on test scores, US

Notes: The estimates of effects of 1-student reductions in class size presented here have been scaled by the authors as detailed in Section 5 and Appendix Table 4 for the sake of reporting estimates in comparable terms. Point estimates are scaled in terms of individual standard deviations. Bars represent the 95% confidence interval.

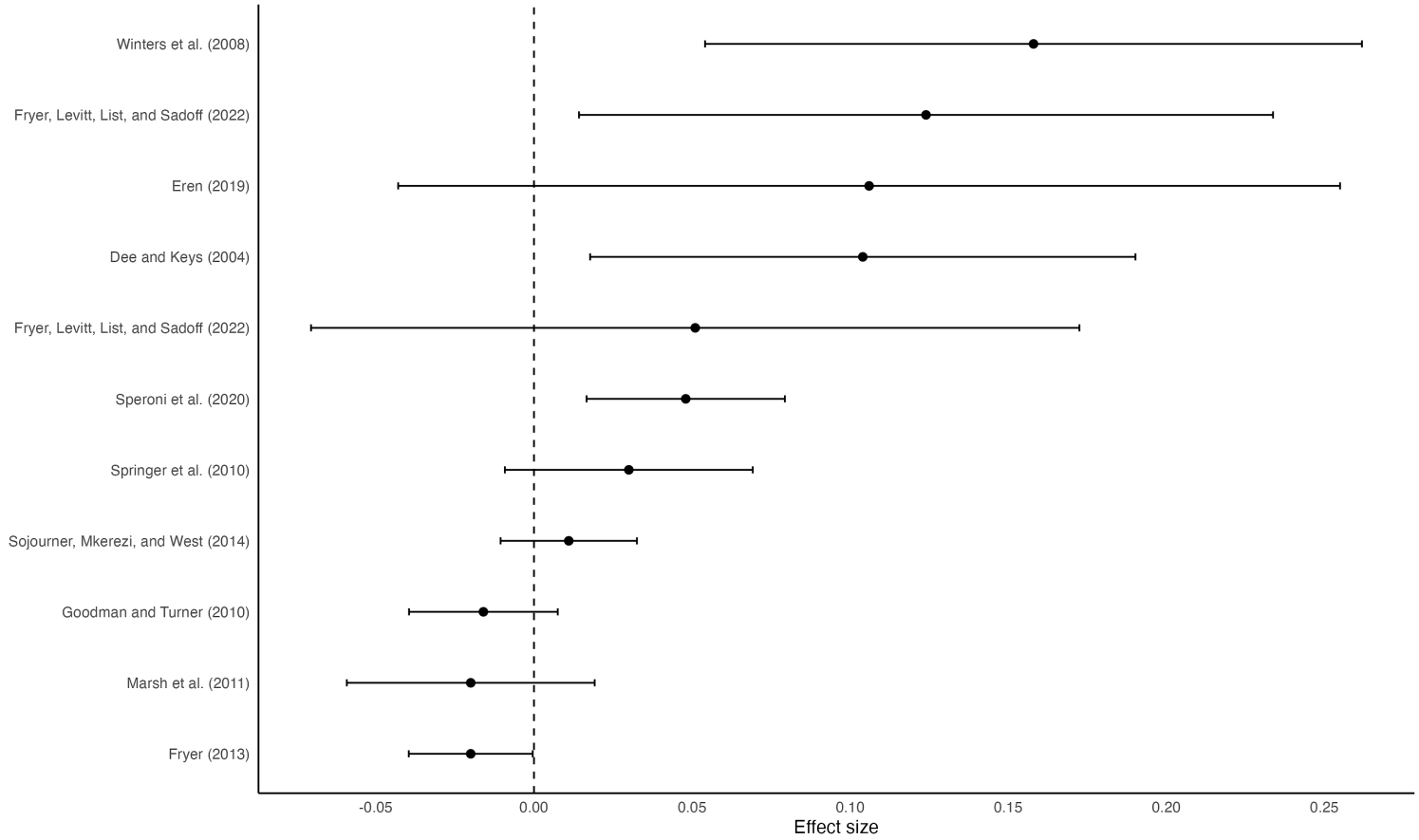

Figure 10: Effects of performance pay on test scores, US

Notes: The estimates of effects of switching to performance-related pay schemes presented here have been scaled by the authors as detailed in Section 5 and Appendix Table 5 for the sake of reporting estimates in comparable terms. Point estimates are scaled in terms of individual standard deviations in test scores, but are not scaled by magnitude of incentives due to data availability issues. Bars represent the 95% confidence interval.