UNIVERSITY OF ARKANSAS | College of Education & Health Professions
Education Reform

WORKING PAPER SERIES

The Participant Effects of Private School Vouchers across the Globe: A Meta-Analytic and Systematic Review

M. Danish Shakeel

Kaitlin P. Anderson

Patrick J. Wolf

May 10, 2016

EDRE Working Paper 2016-07

The Participant Effects of Private School Vouchers across the Globe: A Meta-Analytic and Systematic Review

M. Danish Shakeel

Kaitlin P. Anderson

Patrick J. Wolf

Department of Education Reform

University of Arkansas

201 Graduate Education Building

Fayetteville, AR 72701

Earlier versions of this paper were presented at conferences of the Society for Research on Educational Effectiveness and the Association for Education Finance & Policy. We thank Anna Egalite and Jonathan Mills for details regarding studies included in our analysis. We also thank Mark Lipsey, Phillip Gleason, Gary Ritter and our other colleagues in the Department of Education Reform for comments on earlier drafts. We own all remaining flaws. Corresponding author is M. Danish Shakeel, mdshakee@uark.edu.

Abstract

School voucher programs (a.k.a. opportunity scholarships) are scholarship programs – frequently government funded - that pay for students to attend private schools of their choice. Many private school vouchers programs have been initiated around the world with the goal of increasing the academic performance of students. Voucher programs are often viewed as a way to increase achievement and satisfaction for individual students and families, while at the same time creating competitive pressures that encourage other schools in the area to improve. Countries like Chile and India have developed extensive school voucher programs. While many studies have been conducted on school vouchers, a meta-analysis of the international randomized controlled trials (RCTs) evaluating the achievement effects of vouchers has never been conducted.

This study is a meta-analytic consolidation of the evidence from all RCTs evaluating the participant test score effects of school vouchers internationally. Our search process turned up 9,443 potential studies, 19 of which ultimately were included. These 19 studies represent 11 different voucher programs. A total of 262 effect sizes are included, with a two-stage consolidation of those estimates yielding a total of 44 drawn from the last year of the studies. We have included only math and reading outcomes as other subjects are rarely reported and are difficult to compare across countries. We also differentiate between English and reading outcomes and present English results as a subcomponent of the reading effects to account for the effect of local language in the international context. Our meta-analysis indicates overall positive and statistically significant achievement effects of school vouchers that vary by subject (math or reading), location (US v. non-US), and funding type (public or private). Generally, the impacts are larger (1) for reading than for math, (2) for programs outside the US relative to those within the US, and (3) for publicly-funded programs relative to privately-funded programs.

Keywords: private school voucher, systematic review, meta-analysis, opportunity scholarship, international randomized controlled trials, RCT, participant effects.

1. Background

School choice has emerged as a key demand-side intervention in school reform globally. School vouchers, in particular, are a mechanism by which government resources are provided to families that enable them to attend a private school of their choosing (Wolf 2008a). Strictly speaking, a private school choice initiative is only a "voucher program" if the government funds the program directly out of an appropriation. Other private school choice schemes are funded indirectly, through tax credits provided to businesses or individuals who contribute to nonprofit scholarship-granting organizations. Such arrangements are commonly called tax-credit or opportunity scholarship programs. Other programs, in the U.S. and globally, are funded through private donations and philanthropy, with no specific government tax credit provided. Since tax-credit and privately-funded scholarship programs accomplish the same general purpose as voucher programs – expanding access to private schools of choice for disadvantaged students -- we treat all three types of private school choice programs as functionally equivalent for purposes of this study, although we specify whether individual initiatives are voucher, tax-credit scholarship, or privately-funded programs when discussing them.

Although the origin of the voucher idea is linked to economist Milton Friedman (1955), political philosophers Thomas Paine (1791) and John Stuart Mill (1962 [1869]) supported the theoretical debate about their desirability. The theory of school vouchers is that government should provide funds in support of compulsory education but need not necessarily deliver the schooling itself. Vouchers are a form of government outsourcing. Supporters of vouchers claim that participating students will learn more, either because they will have access to generally higher-quality schools or because their school will be a better match for their particular needs.

Whether or not students benefit from non-governmental organizations providing their education is a fiercely contested empirical question central to the voucher debate (Doolittle & Connors, 2001). For example, Richard Murnane (2005, p. 181) argues:

> Providing families who lack resources with educational choices makes sense. The consequences of attempting to do this through a large-scale voucher…system are unknown. Carefully designed experiments could provide critical knowledge.

Experimental design is critical in the case of evaluating school voucher programs because of concerns about selection bias due to more motivated and able families self-sorting into private schools on their own or through access to a voucher. Fortunately, much of the research on school vouchers in the U.S., and some of the evaluations abroad, has taken the form of random assignment experiments. In this meta-analysis we consolidate the evidence from 19 experimental evaluations of the achievement impacts of private school choice programs in the U.S., India, and Colombia.

2. Private School Choice Programs Around the World

Government or philanthropic efforts to provide greater access to private schools of choice are surprisingly common around the world (e.g. Glenn, De Groof, & Candal 2012a; 2012b; 2012c; 2012d; Wolf & Macedo 2004). Voucher programs generally can be divided into universal and targeted programs. Universal private school choice programs offer government funding of private schooling to all school-age children in a jurisdiction with no eligibility requirements. Universal programs operate in The Netherlands, Belgium, Denmark, Sweden, France and other

European and Commonwealth countries, mainly based on a constitutional right for parents to educate their children within a particular religious, philosophical, or pedagogical tradition (Glenn 1989). A universal school voucher program has operated in Chile since the 1980s (Mizala & Romaguera 2000) and a universal program was enacted for the U.S. state of Nevada in 2015.

Targeted school voucher programs have eligibility requirements that limit private school choice to certain disadvantaged populations of students. Programs funded by philanthropies and limited to low-income students operate in several developing countries in Africa as well as Colombia and regions of India and Pakistan. Many of these programs provide the equivalent of around $200/year to fund schooling at very low-cost private schools operated by education entrepreneurs (Tooley 2009; Dixon 2013). The U.S. was home to 41 targeted private school choice programs as of January 2015, of which 27 were means-tested and 14 were limited to students with disabilities (Frendewey et al. 2015). The vouchers in means-tested U.S. programs range in size from around $1000 to $12,000, as the lower-cost scholarships in that range require families to contribute financially to the cost of tuition. The vouchers for students with disabilities are larger, cover the full cost of educating the child, and in some cases are priced on a sliding scale based on the severity of the child's disability.

In sum, private school choice in a variety of forms exists throughout the U.S. and the wider world. Such programs are increasingly common. The research base on the effectiveness of school voucher programs has been reviewed by multiple scholars over the past eight years but, as seen below, those reviews are inadequate to inform a clear judgement regarding whether students are helped or harmed academically by access to private school choice.

3. A Systematic Review of the Systematic Reviews of Voucher Effectiveness

The ideal meta-analysis is up-to-date, complete, and provides a specific and verifiable determination of the average effect of an intervention on an important outcome (Rossi, Lipsey & Freeman 2004, pp. 324-328; Hedges & Olkin 1985; Hunter & Schmidt 1990). From 2008 through 2015, 10 reviews of the achievement effects of private school choice in the U.S. have been published. None of the 10 satisfy all three criteria for an ideal meta-analysis. Although three of the reviews have been released within the past three years, none of them include the most recent three experimental studies of school vouchers – an important omission since two of those studies include the first estimates of negative achievement effects to come from voucher experiments. None of the reviews included all of the existing school voucher studies within the time-scope and inclusion criteria provided by the authors. Only one review, by Anderson, Guzman, and Ringquist (2013), is a formal meta-analysis that includes overall effect point estimates and confidence levels. The other nine reviews are described by their authors as systematic reviews. The intersection of the reviews of school voucher achievement effects that are up-to-date, complete, and empirically specific is a null set. A current and complete meta-analysis of school vouchers is needed.

First we provide a brief review and critique of the 10 prior reviews of school voucher achievement effects. Table 1 presents information on those reviews that helps us assess their recent vintage and completeness. The 28 empirical achievement evaluations of school voucher programs in the U.S. as of April 2016 appear as rows under the "Study" column, in approximate order from the earliest to the latest. The 10 reviews since 2008 appear as columns, from left to right. The heavy borders in a given column delimit the time scope of the authors' review. Every study that was released during that period should have been included in the review and therefore

should have an "x" in the corresponding table cell. Studies that were categorically excluded based on a legitimate scientific reason, such as because they were merely quasi-experimental, have their cells shaded gray to signify that they are properly disqualified. Naturally, the more recent reviews, on the right side of the table, are much more up-to-date than the less recent reviews, on the left side of the table.

We only provide the final publication for the voucher studies released prior to 2007, since those studies all were completed before the first review in our list. For studies released in 2008 and later, we include annual reports that were part of longitudinal evaluations, since reviewers should have included such interim reports in order to make their review as contemporary as possible. For studies that were published multiple times by similar author groups and which presented the same results based on an identical methodological approach to analyzing the same data (e.g. Peterson et al. 2003 & Mayer et al. 2002), we only classify them as a single study, both in this review of the reviews and in our meta-analysis to follow.

[Table 1 about here]

The first review of school voucher achievement effects released since 2008 was published by The Great Lakes Center for Education Research & Practice (Miron, Evergreen & Urschel 2008), a think-tank generally viewed as hostile to market-based education reforms such as vouchers. The review included 12 of the 15 empirical voucher studies that existed at that time, omitting Jay Greene's (2000) experimental evaluation of the Charlotte privately-funded scholarship program and a similar experimental evaluation of the New York City privately-funded scholarship program conducted by a group of prominent statisticians (Barnard et al.

2003) as well as the lesser-known Bettinger & Slonim (2006) experimental evaluation of a privately-funded scholarship program in Toledo. Two of the excluded studies reported positive effects of vouchers on student achievement in both reading and math (Greene 2000) or only in math and only for African American students (Barnard et al. 2003) while the third study identified no significant voucher impacts from a small analytic sample (Bettinger & Slonim 2006). The Great Lakes Center review concluded that, "voucher studies, generally of high quality, indicate a slightly positive impact, particularly for African American Students." (Miron, Evergreen & Urschel 2008, p. 1).

The second and third reviews both were published in the same law review journal in the wake of an academic conference on school vouchers (Lubienski & Weitzel 2008; Wolf 2008b). The Lubienski and Weitzel (2008) review focused on the purported political motivations of voucher evaluators but included a section that reviewed the existing research literature on school vouchers. That review excluded nearly 40% of the empirical studies published prior to the submission of the final manuscript, four of which reported at least some positive effects of vouchers and two of which reported only null findings. Lubienski and Weitzel (2008, p. 462) concluded: "positive academic outcomes stemming from voucher programs are modest at best, do not extend to most groups, and certainly do not rise to the level anticipated by the early optimistic assumptions advancing such programs." Wolf (2008b) limited his review of the evidence to 10 of the 11 experimental voucher evaluations that existed at the time. Like all the other reviewers, he failed to include the Bettinger & Slonim (2006) study of Toledo. He concluded (p. 466): "We know, through the assistance of a substantial body of rigorous experimental studies, that the effect of vouchers on student achievement tends to be positive;

however, achievement impacts are not statistically significant for all students in all studies and they tend to require several years to materialize."

The final of the four voucher reviews of 2008 was a National Center for the Study of the Privatization of Education working paper by Rouse & Barrow (2008) later published in the *Annual Review of Economics* (Rouse & Barrow 2009). Although they did not state that their review was limited only to experimental studies, they lauded the rigor of experiments, which comprised seven of the eight studies they reviewed. They committed both Type I (including a non-experimental study) and Type II (excluding several experimental studies) errors in their sample of studies, inexplicably including a single quasi-experimental voucher evaluation of the Cleveland program (Belfield 2006) while excluding all the other quasi-experimental voucher studies as well as more rigorous and more positive experimental evaluations of voucher programs in Milwaukee (Greene, Peterson & Du 1999), Charlotte (Greene 2000; Cowen 2008) and New York City (Barnard et al. 2003). Rouse and Barrow (2008, abstract) concluded: "The best research to date finds relatively small achievement gains for students offered education vouchers, most of which are not statistically different from zero."

Andrew Coulson (2009) produced a vote-counting meta-analysis of achievement effects for all quantitative studies that compare the private to the public provision of education. His sub-sample of 11 voucher studies, which included 65% of the studies then extant, included just two scored "1" (Greene, Peterson & Du 1999; Rouse 1998), because the dominant finding was overall positive effects, and one scored "-1" (Belfield 2006), because the dominant finding was overall negative effects. The other eight studies all were scored "0" because the overall effect of school vouchers on achievement was not statistically significant in those evaluations. Coulson did leave out of his review four voucher studies that reported positive test score effects and two

studies that reported no significant effect, so a proper vote count from his study would have been a net score of 5 in a range between possible scores of -17 to 17.

The first voucher review of 2011 was published by Usher et al. (2011) at the Center on Education Policy, a DC think tank generally viewed as opposing market-based reforms such as school vouchers. The authors excluded all studies prior to 2000, since they had published a similar review of the existing voucher literature that year. They also excluded all studies of privately-funded scholarship programs, a decision that removed from their sample many of the most rigorous and positive voucher evaluations. The CEP review is the only one of the 10 systematic reviews to exclude Rouse (1998), the three studies by Peterson et al. (2003), and Krueger & Zhu (2004) from consideration. Its purportedly comprehensive review only included 53% of the school voucher evaluations to that point. The study concluded (Usher et al., 2011, p. 9): "Achievement gains for voucher students are similar to those of their public school peers."

Forster (2011; 2013) authored two reviews of the research evidence on school vouchers published by the pro-voucher philanthropy The Friedman Foundation for Educational Choice. Forster limited his review to the results from "gold standard" experimental studies. His 2011 review captured 13 of the 15 experimental studies that existed at that time and his 2013 update included 14 of 15, only missing Bettinger & Slonim (2006). Forster classified a study as "positive" regarding the achievement effects of school vouchers if it reported any statistically significant positive impacts and no statistically significant negative impacts. He classified a study as "neutral" if all of the findings it reported were not statistically significant. By Forster's (2013) vote count, 13 rigorous experimental evaluations were positive regarding the effects of school vouchers on student achievement whereas just one study (Krueger & Zhu 2004) was neutral and none were negative.

The only statistical meta-analysis of school voucher achievement effects was published in 2013 by Anderson, Guzman & Ringquist. It appeared as a chapter in a textbook on using meta-analysis to guide public administration and policy. The researchers sought to include every statistical evaluation of private school choice programs in the U.S. in their sample, regardless of rigor, but only actually captured 68% of the studies then extant.[1] Of the eight studies missed by this meta-analysis, three of them reported positive voucher effects and five found no impacts. The researchers counted every overall and subgroup estimate of voucher impact from every study as a separate observation, analyzing 611 effect estimates in total. Over one-third of their estimates came from a single data-base that informed all of the studies of the Cleveland Scholarship and Tutoring Program (Greene et al. 1998; Metcalf 2003; Plucker et al. 2006; Belfield 2006). The Cleveland program itself has certain peculiarities, including that it provides the lowest-value voucher of any of the government-run voucher programs in the U.S. The Cleveland evaluation itself was non-experimental, with a matched-sample comparison group, suffered by high levels of sample attrition that were disproportionate to the comparison group, and included a lot of alternative estimates of impacts in part because the data were of such poor quality. Given that the meta-analytic approach of the authors implicitly weighted the weakest of the voucher evaluations much more heavily than the stronger studies, it is not surprising that the meta-analysts concluded (Anderson, Guzman, & Ringquist 2013, p. 336): "vouchers have had a positive and significant but substantively trivial effect on student academic achievement."[2]

[1] The senior researcher on the project, Evan Ringquist, died shortly after the book was published, after a long battle with cancer. It is likely that his health explained at least some of the notable study omissions after 2010.

[2] The overall average effect of school vouchers on student achievement was calculated by the authors to be +.03 standard deviations (SD), leading to their conclusion that voucher effects are positive but trivial in size. Later in the study, using meta-regression, they conclude (p. 346) that "Design characteristics and the quality of original studies exert the largest influence over effect sizes…lower-quality studies estimate smaller average effect sizes." Their d-based estimate of the average effect of vouchers on student

The most recent review of the school voucher literature was published in 2015 by Epple, Romano & Urquiola. It was described by the authors as a review of the economics literature on school vouchers. The research base on school vouchers is highly inter-disciplinary, with substantial contributions from political scientists, statisticians, sociologists, and education researchers in addition to economists. By informally restricting their review to voucher evaluations either by or easily known to economists, the authors missed over half of the empirical studies to date, including seven of the eight studies released from 2006 to 2009. In this least complete of voucher reviews, the authors concluded (Epple, Romano & Urquiola 2015, abstract): "many studies find insignificant effects of voucher on educational outcomes; however, multiple positive findings support continued exploration."

Our review of the reviews of the school voucher research literature makes a strong case for our meta-analysis. The previous reviews varied greatly in their methodology, search criteria, effectiveness of search, and conclusions. Based on these reviews, school vouchers have no effect on student achievement (Usher et al., 2011), consistently improve achievement (Forster 2011; 2013), or produce some mix of positive effects and no significant effects that is either encouraging (Miron, Evergreen & Urschel 2008; Wolf 2008; Rouse & Barrow 2008; Coulson 2009; Epple, Romano & Urquiola 2015) or disappointing (Lubienski & Weitzel 2008; Anderson, Guzman & Ringquist 2013). Most of the individual studies had analytic samples of less than 1000 students in the final evaluation year and therefore were at best modestly powered to detect voucher effects with a reasonable level of confidence. The many findings of "no significant effects" from these voucher studies could be due to a low signal/noise ratio or because private school vouchers truly have no effect on student achievement. Only the Coulson (2009) and

achievement, when limited to the experimental evaluations of voucher programs, is nearly .11 SD, nearly four times the magnitude of their overall estimate (p. 348).

Epple, Romano & Urquiola (2015) reviews included studies from outside the U.S. and none of the reviews included recent studies from 2015 and 2016. Given the lack of any contemporary, complete meta-analysis of the effect of private school vouchers on student achievement around the world, we think the motivation for our study is especially strong.

4. METHODS

4-A. Search Strategy

For this meta-analysis we identified publications from computer and networked searches through a variety of sources. To begin, we explain the two stages of our search. Then, we outline our specific search strategies and selection criteria. Finally, we explain the methods used to determine whether to include or exclude sources, how we extracted data and finally coded the information for the meta-analysis.

Our initial search focused on only the studies published since 2005 or later, but due to a lack of RCT studies identified during this process, we added a second search, including all years, but narrowing the search criteria to only include studies that included text related to randomization.

We focused especially on identifying experimental (a.k.a. RCT) voucher studies, and eventually decided to limit our meta-analysis to them, for several reasons. First, RCTs are the "gold standard" of program evaluation in terms of assessing causal relationships (e.g. Boruch & Mosteller, 2002; Rossi, Lipsey & Freeman 2005). The random assignment of subjects that is a defining feature of RCTs creates a treatment group (in this case, those receiving the offer of a voucher) and a control group (those who did not receive the offer of a voucher) that are similar to each other in expectation regarding all measurable and unmeasurable characteristics. This

similarity achieved by RCTs is especially important when evaluating private school choice programs, since families who self-select into private schools are widely expected to be different from other families in unmeasurable ways that subsequently affect student achievement levels and gains. In RCTs, access to private schooling through a voucher is random, therefore solving the selection bias problem at least in theory. Often, quasi-experimental methods that attempt to account for this selection bias fall short (Betts et al., 2010).

Second, we know from the previous statistical meta-analysis of school voucher achievement effects that the conclusion one draws about the efficacy of vouchers is heavily influenced by which body of studies one reviews. The quasi-experimental studies tend to produce smaller voucher effect sizes and fewer statistically significant results, arguably because weaknesses in the research design and comparison groups bias the impact estimates towards 0. If one has to believe either the results from RCTs or the results from non-RCTs regarding the effects of a given intervention, because the results differ substantially, then one should believe the results from RCTs because they have much stronger internal validity.

Third, we expected that there would be a sufficient number of voucher effect estimates from RCTs in our sample to produce a reliable estimate of voucher impacts. Since the geographical scope of our search was global, and the final temporal scope of our search was unrestricted, we thought that we would identify a critical mass of voucher achievement studies even restricting our sample to gold standard experiments. The fact that we ended up with 19 studies that provided over 100 effect estimates confirmed the wisdom of our approach.

Whenever one can restrict one's evidence to findings from gold standard RCTs, one should do so. We can, so we do.

The study selection was based on systematic search procedures. Keywords and phrases were chosen to be as inclusive as possible for our preliminary search. The team used EBSCO, JSTOR, and ProQuest databases through the library of the University of Arkansas. In addition, we used a Google Scholar search and other additional websites to further identify any sources missed in these three databases. Lastly, we utilized subject matter experts in the field and snowballing techniques to find additional relevant studies. All of our searches identified 9,443 articles that could be relevant to our meta-analysis.

The search criteria were as follows:

Initial Search: 2005 or later

EBSCO Search 1

Search terms: school voucher* OR education* voucher*

Time period: 2005 or later

Types of sources included: Academic Journals, Journals, and Reports

Total number of results: 765

EBSCO Search 2

Search terms: opportunity scholarship

Time period: 2005 or later

Types of sources included: Academic Journals, Journals, and Reports

Total number of results: 48

JSTOR Search 1

Search terms: voucher* AND education* or school AND research AND experiment* or "randomized controlled trial"

Time period: 2005 or later

Language: English

Included only Articles related to: Business and Economics, Economics, Education, Political Science, Public Policy & Administration, Social Sciences

Total number of results: 853 search results

JSTOR Search 2

Search terms: "opportunity scholarship"

Time period: 2005 or later

Language: English

Included only Articles related to: Business and Economics, Economics, Education, Political Science, Public Policy & Administration, Social Sciences.

Total number of results: 30 search results

ProQuest Search 1

Search terms: all(voucher) AND all(school*) AND all(research*)

Time period: 2005 or later

Excludes: Wire Feeds, Magazines, and Newspapers

Total number of results: 603 results

ProQuest Search 2

Search terms: all("opportunity scholarship")

Time period: 2005 or later

Excludes: Wire Feeds, Magazines, and Newspapers

Total number of results: 122 results

The searches of the three library databases (EBSCO, JSTOR, and ProQuest) resulted in a total of 1,934 unique papers, after removing duplicates.

Secondary Search: All RCTs (including prior to 2005)

Since RCTs or experiments are especially prized as education evaluations, we decided to extend our meta-analysis to any RCTs we could find on the topic, regardless of when they were conducted or published. In order to find these, a secondary search was conducted.

EBSCO Search 3 (for all RCTs)

Search terms: school voucher* OR education* voucher* AND AB: random*

Time period: No restriction

Types of sources included: Academic Journals, Journals, and Reports

Total number of results: 85

Note: AB: random means that the abstract had to include a stem of the word random**

EBSCO Search 4 (for all RCTs)

Search terms: opportunity scholarship AND AB: random*

Time period: No restriction

Types of sources included: Academic Journals, Journals, and Reports

Total number of results: 9

Note: AB: random means that the abstract had to include a stem of the word random**

JSTOR Search 3

Search terms: voucher* AND education* or school AND ab(random*)

Time period: No restriction

Language: English

Included only Articles related to: Business and Economics, Economics, Education, Political Science, Public Policy & Administration, Social Sciences

Total number of results: 116 search results

JSTOR Search 4

Search terms: "opportunity scholarship" AND ab(random*)

Time period: No restriction

Language: English

Included only Articles related to: Business and Economics, Economics, Education, Political Science, Public Policy & Administration, Social Sciences.

Total number of results: 2 search results

ProQuest Search 3

Search terms: all(voucher) AND all(school*) AND all(research*) AND ab(random*)

Time period: No restriction

Excludes: Wire Feeds, Magazines, and Newspapers

Total number of results: 95 results

Note: ab(random) means that the abstract had to include a stem of the word random**

ProQuest Search 4

Search terms: all("opportunity scholarship") AND ab(random*)

Time period: No restriction

Excludes: Wire Feeds, Magazines, and Newspapers

Total number of results: 9 results

Note: ab(random) means that the abstract had to include a stem of the word random.*

This secondary search of the three library databases (EBSCO, JSTOR, and ProQuest) resulted in a total of 269 additional unique papers, after removing duplicates.

Google Scholar and Other Website Searches

In addition to the three main library databases, we searched a variety of other sources. First, using the first search criteria, we searched Google Scholar for articles from 2005 or later using the search terms "school voucher" OR "voucher school" to find the maximum number of results. The search returned approximately 4,000 results including patents and citations. Other places we searched, due to their interest in school vouchers, were the websites of the National Bureau of Economic Research (NBER), University of Chile, Uppsala University in Sweden, and the Poverty Action Lab at MIT.

Using the second search criteria in Google Scholar: (("opportunity scholarship" OR "education* voucher*" OR "school voucher*") AND random*), we found 2,570 results including

citations. Apart from importing the references in Refworks, we also did individual Google Scholar searches of the imported references whose titles did not end up in Refworks. We also added three studies found through a networked search, that were not published at the time of our systematic review searches (Abdulkadiroglu et al. 2015; Wolf, Egalite, & Dixon 2015; Mills & Wolf, 2016).

4-B. Selection Process

These additional-non-library sources were combined and then 6,549 were excluded based on title and/or abstract reviews. Each of these sources were reviewed by two separate team members based on their title and abstract in order to determine whether they should progress to the next stage, in which we reviewed the full articles. In some cases, there was a disagreement between the researchers on whether to include or exclude a particular study, so the two team members came to a conclusion together. Unless there was a clear reason to exclude the paper, it was kept until the full article review round, when more information would be available to judge.

As mentioned previously, we conducted two rounds or searches, one for all empirical voucher studies since 2005 and a secondary search for all RCTs ever conducted on the topic. After both of these searches, we determined which articles to include based on several criteria. To be included in the meta-analysis, the studies had to focus on the participant effects of private school vouchers and measure quantitative test score outcomes in either math or reading. Studies dealing with other impacts of vouchers such as competitive effects or fiscal impacts were

excluded. We did not include graduation rate, college attainment, or civic values outcomes in the current study. We only included studies published in English or with English translations.[3]

After our title/abstract review, 148 sources remained from the Google and snowball search along with 128 sources from the library searches. Two members of our team reviewed each of these 276 sources in their entirety in order to determine if they met our inclusion criteria. In some cases the researchers initially disagreed on the inclusion decision. They then met to discuss the case and came to a consensus decision. Common reasons for exclusion were that the studies were theoretical discussions or opinion pieces without rigorous evidence, they focused on other issues related to school vouchers such as competitive effects or fiscal impacts rather than participant effects, they were merely quasi-experimental,[4] or they were RCTs that did not report outcomes on math, reading/English. Our full-article review process resulted in 16 studies remaining in the sample.

Last, we conducted a final network search based on matching our list of potential sources with earlier publications on school vouchers internationally and current voucher evaluations compiled by Patrick J. Wolf, a co-author of this study. This final review resulted in three additional articles added to the sample – two of the recently implemented Louisiana voucher program and one of a philanthropic voucher program in Delhi, India. In the end, 19 RCT studies made the final cut. Appendix B contains the details regarding the studies that were identified and eliminated at each stage. In Table 2 we summarize the studies, presenting attrition rates in terms of both sample attrition (the percent of study participants who are not observed in any particular

[3] We did not search dissertation or master's thesis databases because we expect that any experimental evaluation of a school voucher program that is the subject of an original thesis or dissertation will be sufficiently important that it also will be released as a study report or journal publication.
[4] A surprising number of education evaluations are described as "experimental" via keywords or in their abstracts but, upon a closer reading, actually do not create their comparison groups via random assignment and therefore are merely quasi-experimental.

year) and program attrition (the percent of students offered a voucher who do not use the voucher in any particular year).

[Table 2 about here]

The global scope of our search was especially fruitful in identifying rigorous school voucher evaluations that have been omitted from many previous reviews of the research evidence on private school choice. Two different studies of a large voucher program in Bogota, Colombia (Angrist et al. 2002; 2006), and two studies of different programs in separate regions of India (Muralidharan & Sundararan 2015; Wolf, Egalite, & Dixon 2015) were uncovered through our search. Our combined computerized and networked search also identified an RCT of a small privately-funded voucher program (in Toledo) that had never before been included in a review of voucher research (Bettinger & Slonim 2006). Finally, we were able to include three major evaluations of recent vintage that also have never informed a school voucher review (Bitler et al. 2015; Abdulkadiroglu et al. 2015; Mills & Wolf 2016). This represents a new look at a much more comprehensive body of rigorous research on private school vouchers than ever before.

Many of the published reports of experimental evaluations of school voucher programs are nested in various ways that affect how much independent information they contribute to a meta-analysis. For example, at least six different research teams have published more than two dozen reports or articles analyzing the experimental data from the New York City Children's Scholarship Fund evaluation, 1998-2002. Including all 24 or so of those reports would generate a substantial amount of spatial auto-correlation in our analysis due to "double-counting" of

effects. We decided that any publications of the same results, using the same methodology, by essentially the same research team were essentially a single study. Any variation on that, such as publication of different results, using the same methodology, by a different research team (e.g. a failed replication), represented a different study even though it drew upon the same data. That determination reduced the number of New York City studies to five. We then extracted most of the data from the final publication in the series, unless an earlier publication contained more complete data, and supplemented those data with additional details from other studies in the "nest" as needed. In essence, a "study" in our meta-analysis is the final and most complete presentation of a specific set of findings from a specific research team using a particular analytic method.

4-C. Programs Included in the Meta-Analysis

The 19 RCT studies identified by our search represent 11 separate school voucher programs (Table 3). Six programs – in Andhra Pradesh and Delhi, India; Toledo and Dayton, Ohio; and the DC WSF and OSP -- were subject to a single experimental evaluation. Four programs – in Charlotte, NC; Louisiana; Milwaukee, WI; and Bogota, Colombia -- were the focus of both an original experimental study and one replication study. The New York City program was the subject of five different experimental analyses.

In Table 3, each program is categorized as either privately versus publicly funded (where public funding programs are defined as those with *any* amount of public funding, and privately funded programs as those that are *exclusively* privately funded, through development or philanthropic funds), and as either fully or partially funded vouchers. In general, the fully funded

vouchers are publicly funded and the partially funded vouchers are privately funded. Funding for the programs in India and Colombia, whether "full" or "partial," was extremely low, ranging from about $117 in India to $190 in Colombia, in nominal U.S. dollars (Wolf, Egalite & Dixon 2015; Angrist et al. 2002). The "fully funded" programs in the U.S. provided vouchers with maximum values that ranged from around $5,000 in Louisiana to $7,500 in DC (Mills & Wolf 2016; Wolf et al. 2013). Partially funded programs in the U.S. generally provided about $2,000 in tuition support to families (Peterson et al. 2013). Regardless of jurisdiction and full or partial funding, the maximum voucher values for all of the programs in this meta-analysis represented less than half of the amount that was being spent per-pupil on students in area public schools.

All of the programs are targeted to low-income students through either income limits or program location, but usually both. The voucher initiatives in India and Colombia serve students living in abject poverty (Muralidharan & Sundararaman 2015; Wolf, Egalite & Dixon 2015; Angrist et al. 2002; Tooley 2009). The U.S. programs all are limited to students with family incomes near or below the cut-off for the federal lunch program. All of the U.S. voucher initiatives in this meta-analysis are limited to cities except for the statewide Louisiana Scholarship Program. The overwhelming majority of voucher participants in the U.S. are either African American or Hispanic. This is a study of the achievement effects of low-cost private school vouchers on low-income inner-city children.

The private schools participating in these voucher programs tend to charge modest tuition and have experience serving disadvantaged student populations. Religious schools in general, and Catholic schools in particular, are the main participants in voucher programs in the U.S. In the D.C. Opportunity Scholarship Program, for example, 80% of the participating students attended a religious school with their voucher and 53% of them specifically attended a Catholic

school (Wolf et al. 2013, p. 257). Across programs, the private schools serving students with vouchers tend to be "no frills" with modest school facilities and few special programs for differentiating instruction to students (e.g. Wolf et al. 2013; Dixon 2013). They tend to provide a consistent educational experience to all students focused on academic fundamentals and character development.

The counterfactual condition for control group students varied across the programs. In India and Colombia, almost all of the students who lost the voucher lotteries attended government-run schools in their neighborhoods. In India especially, public schools are much better resourced than low-cost private schools but are plagued by teacher absenteeism rates of around 30% (Probe Team 1999). Few public schools in developing countries arrange for substitute teachers. In cases where public school teachers fail to show up for work, the children are on their own.

In the U.S. voucher programs in our meta-analysis, students who lost the voucher lotteries often found other ways to access school choices. In the experimental evaluation in Dayton, Ohio, 18% of the control group students enrolled in a private school even without the assistance of a voucher (Howell & Peterson 2006, p. 44). In the DC OSP study, 12% of the students that lost the lottery subsequently enrolled in a private school and 35% attended an independent public charter school, leaving just 53% of the control group students in traditional public schools (Wolf et al. 2013, p. 257). In Louisiana, only 6% of the control group students enrolled in a private school after losing the lottery but 29% of them attended a public school of choice, leaving just 65% in a traditional public school (Mills & Wolf 2016, p. 21). The New York City program demonstrated the clearest treatment-control contrast in type of school attended, as only 4% of the students that lost the lottery attended a private school on their own

and public charter schools were uncommon in the city during the study period so almost all of the control group was in traditional public schools (Howell & Peterson 2006, p. 44). In the experimental studies included in this meta-analysis, students remained in the control group and their outcomes counted towards the control group average for the Intent-to-Treat (ITT) impact estimates even if they attended a private school. The rates at which control-group students crossed over to private schooling factored into the Treatment-on-Treated (TOT) effect calculations, however.

[Table 3 about here]

4-D. Data Extraction

The remaining nineteen studies were coded in Microsoft Excel for details on author, publication year, location, funding type (public/private), years of evaluation, duration of study, grades analyzed, outcome (reading/English or math), size of treatment and control group and overall sample size. Finally, some studies had multiple evaluation years for the same program. Each evaluation year, type of impact estimate (Intent-to-Treat [ITT] or Treatment-on-Treated [TOT]), and subject was treated as a separate observation in the database. A study that reported results in each of three years, in both reading and math, that included both ITT and TOT estimates, contributed 12 observations to the database (3 x 2 x 2), but ITT and TOT estimates, math and reading estimates, and estimates from the same study over different years were never combined. The 12 observations that a given study might produce would only be analyzed within a specific meta-analytic estimate of effect, such as the ITT estimate of the voucher effect in math in Year 2 after random assignment. When the authors provided results from multiple estimation models or

from robustness checks, we only extracted the estimates from the "most preferred model" as signaled by the authors or the final model if no preference was given.

All the information was extracted using a predesigned (but modifiable) coding form in Excel. The extracted data filled 262 rows of an Excel spreadsheet, meaning a total of 262 distinct effect estimates informed our meta-analysis. A total of 70 of the estimates are ITT effects in reading, 62 are ITT effects in math, 69 are TOT effects in reading, and 61 are TOT effects in math. The extraction process was performed independently by at least two team members so they could match their results and minimize human error. As some studies did not report their findings in detail, we made necessary assumptions to derive accurate sample sizes for the treatment and control groups. See Appendix A for details on the assumptions made for each study in which a key data point had to be calculated because it was not provided in the source.

4-E. Data Synthesis

The meta-analysis of the RCTs essentially creates an overall effect size by combining the effect sizes extracted from each study. Effect sizes were analyzed separately for math and reading/ English outcomes. Both intent to treat (ITT) and treatment on the treated (TOT) effects are calculated, when possible. The overall effect size in the meta-analysis is based on a weighted average of the individual effect sizes, across years, extracted from the studies. Each observation's weight was set as the inverse of the variance around the effect size, so effects that were estimated more precisely were weighted more heavily. The effect size and standard errors were extracted directly from the source if available. If these numbers were not reported, they were calculated by the team using the data that were available and the formulas in Appendix C.

Effect sizes give the size treatment impact in standard deviation units, so it is a measure that can be averaged over several studies. The standard errors on these effect sizes indicate a measure of variance and are used to create a confidence interval around the point estimate of the effect size.

One of the benefits of the meta-analysis is that it combines results from several studies, which can often individually have small sample sizes and low precision. For the meta-analysis, we used MS Excel and STATA for the final estimates to double check for estimation errors. We calculated the pooled standard deviation and effect size using Hedges' *g*. We also calculated the standard error for the effect size. Lastly, the grand effect size and lower and upper bound of the overall 95% confidence interval were calculated. The nineteen RCT studies that we included in the meta-analysis primarily measured math and reading outcomes. Only one study (Bettinger & Slonim, 2006) had only math test outcomes.

The entire analysis was performed in two steps. In the first step, we estimate an overall Intent to treat (ITT) and Treatment on treated (TOT) effect for each year of the outcomes available for Reading/ English and Math for each program by combining estimates reported across different studies for the same program in the same year. This "mini-meta-analysis" of findings by site-year reduced the total number of effect size estimates to 98: a total of 23 estimates for ITT Reading and English; 24 estimates for ITT Math; 25 estimates for TOT Reading and English; and 26 estimates for TOT Math. At this stage, a fixed effects meta-analysis was conducted, since for the set of students pertaining to a particular program, the data was essentially the same, and therefore we assume the true effect is the same in all studies (Borenstein, Hedges, & Rothstein, 2007).

In the second step, we estimate overall voucher effects using a fixed effects meta-analysis. Despite these estimates coming from different studies, there were too few studies in

some cases to justify the use of random effects. Use of random effects would not result in precise estimates as the between-studies variance cannot be estimated with precision. In such a case, fixed effects is the only viable option (Borenstein, Hedges, & Rothstein, 2007). The overall ITT and TOT (Reading/ English and Math) voucher effects are analyzed based on geography (US vs. International and an overall global effect), funding type (publicly funded vs. privately funded programs) and years of treatment (one year, two year, three year and four or more years of being in the treatment). The analysis for years of treatment uses all the 98 effect size estimates (which themselves represent a consolidation of the 262 extracted estimates) and all other analyses are based on the 44 effect size estimates for the last year covered by each study (11 estimates for ITT Reading and English, 10 estimates for ITT Math, 12 estimates for TOT Reading and English, and 11 estimates for TOT Math.

5. RESULTS

We discuss the results of the school voucher RCT meta-analysis in terms of overall average treatment effects and with a specific focus on outcomes by type (ITT or TOT), subject (reading or math), location (US or non-US), and type of funding (public versus private). In addition, we provide results by years of treatment (1, 2, 3, and 4 or more). Each of these impacts is calculated using Hedge's g, and we include a 95% confidence interval around each estimate.

First, we present the global results for reading and math (ITT and TOT). English results will also be presented as a subcomponent of the reading effects for countries where English is not the native language but is taught in schools. For each of these effects, we also compare US and non-US programs. Next, we split the findings into public versus private (again noting ITT and TOT effects). Finally, we will present the results by year.

To present our results we provide various forest plots, which show the effect size and confidence interval for each study, for the US and non-US components, and overall. Individual studies are represented by the box and whisker plots, where the size of the box represents the relative weighting of that study and the length of the whiskers represents the confidence interval. Any confidence interval that crosses zero signals that an effect is not statistically significant. The diamonds represent composite effects across all observations.

5-A. Overall Impacts

Figure 1 presents the global ITT reading impacts. The offer of a voucher has a statistically significant and positive impact of about 0.17 standard deviations [95% CI: 0.15, 0.20]. This overall effect is driven by four programs that had positive effects with 95% confidence (one in the US and three outside of the US). Comparing the six US and three non-US programs that we had reading impacts for, we see that the US programs had an overall effect that was barely a null effect, but tended towards a positive effect [95% CI: -0.00, 0.08]. On the other hand, the programs outside of the US had a more definitive positive impact on reading scores of 0.24 standard deviations [95% CI: 0.21, 0.27].

[Figure 1 about here]

Looking specifically at English impacts in Figure 2, we see a positive, yet somewhat smaller impact of 0.08 standard deviations [95% CI: 0.04, 0.11]. This impact was driven by three programs with significantly positive effects (one in the US and two outside of the US). The US effects in English are the same as the reading effects, because within these programs, tests were

not administered within the US in any other languages.[5] The overall effect of programs outside the US was smaller in English (0.13 standard deviations) than in all languages (0.24 standard deviations).

[Figure 2 about here]

Figures 3 and 4 present the same types of forest plots for the TOT effects for reading globally and English globally. In addition, composites of the US and non-US effects are provided. As expected, the TOT effects are at least as large as the ITT effects (0.27 standard deviations in reading and 0.08 standard deviations in English). The TOT effect in reading (including all languages) was primarily driven by a very large effect in the PACES program in Bogota, Colombia (1.4 standard deviations). These TOT effects represent the average treatment for a voucher user, and are generally larger than the ITT effects due to non-compliance.

[Figure 3 about here]

[Figure 4 about here]

Figure 5 presents the ITT results for math globally. The offer of a voucher has a positive impact of 0.11 standard deviations on student math scores [95% CI: 0.08, 0.14]. This effect is driven primarily by two programs with positive effects (one in the US and one outside of the

[5] The reading exams were administered in Spanish for the Angrist et al. (2002; 2006) evaluations of the Colombia Program, English, Telugu and Hindi for the Muralidharan & Sundararaman (2015) evaluation of a program in India, and English and Hindi for the Wolf et al. (2015) evaluation of the Delhi program.

US). In this case, both effects are positive and statistically significant for both the US programs (0.07 standard deviations, [95% CI: 0.02, 0.11]) and the non-US programs (0.15 standard deviations, [95% CI: 011, 0.19]).

[Figure 5 about here]

The global TOT effects in math are somewhat larger (see Figure 6). Using a voucher improved math scores by 0.15 standard deviations, on average [95% CI: 0.12, 0.18]. The US programs, overall, had a TOT effect that was not statistically different from zero [95% CI: -0.05, 0.04]. The programs outside of the US had a positive TOT effect of about 0.33 standard deviations [95% CI: 0.29, 0.37]. As expected, the TOT effects are expected to be larger than the ITT effects, in general.

[Figure 6 about here]

The overall results so far indicate that school vouchers have positive effects in both reading and math, but that these impacts are largest in programs outside of the US. Next, we look at the programs globally, and separate the effects by funding type (private or public). For the purposes of this distinction, we define publicly funded programs as those with *any* amount of public funding, and privately funded programs as those that are *exclusively* privately funded, through development or philanthropic funds.

5-B. By Funding Type

Figure 7 presents the ITT results in reading, by funding type. Both the publicly- and privately-funded voucher programs have positive effects on reading, overall. Privately-funded programs improve the test scores of voucher winners by 0.09 standard deviations, on average, and publicly-funded programs improve the test scores of voucher winners by 0.45 standard deviations, on average. Again, this is driven primarily by one large positive impact of the PACES program in Bogota, Colombia.

[Figure 7 about here]

The corresponding TOT results in reading, by funding type, are presented in Figure 8. These results are even larger due to scaling up by the usage rate. Voucher users in privately-funded programs have positive impacts in reading of 0.15 standard deviations, but voucher users in publicly-funded programs experience much larger reading impacts (0.69 standard deviations).

[Figure 8 about here]

Figure 9 presents the ITT results in math, by funding type. Privately-funded programs do not affect math scores for those offered a voucher [95% CI: -0.01, 0.06]. Publicly-funded programs, on the other hand, have a positive ITT effect of 0.29 standard deviations [95% CI: 0.24, 0.35]. The TOT results in math for privately-funded programs are also null (see Figure 10), but the TOT impacts for publicly-funded programs are an increase of 0.36 standard deviations [95% CI: 0.31, 0.41].

[Figure 9 about here]

[Figure 10 about here]

5-C. By Years of Treatment

The last comparison of effects we present in this meta-analysis is the effects on reading and math by years of treatment. If there is a cumulative positive effect of voucher treatment over time, we would expect impacts to increase with the number of years of access to or usage of the voucher. These results are presented for 8 programs with one year effects (only 7 programs with ITT effects), 8 programs with two year effects (only 7 programs with ITT effects), 6 programs with three year effects in math (5 in reading), and four programs with effects of four or more years of treatment.

Figure 11 presents the ITT reading impacts by years of treatment. The offer of a voucher had a null effect on students after one year, small impacts on students after two or three years (0.04 standard deviations and 0.05 standard deviations, respectively), and a somewhat larger impact after four or more years (0.24 standard deviations, [95% CI: 0.21, 0.28]). Generally, we do see a positive trend in ITT reading effects over time.

[Figure 11 about here]

Figure 12 shows a forest plot for TOT reading impacts, by years of treatment. There was a null effect of one year of treatment, small effects for two and three years of treatment (0.08

standard deviations and 0.06 standard deviations, respectively), and a large effect (0.54 standard deviation) for four or more years of treatment. Again, as expected, the treatment effects tend to increase with time of exposure.

[Figure 12 about here]

Turning to the ITT math results (Figure 13), we see positive impacts for one year, three years, and four or more years of treatment, but null effects for two years of treatment. Students offered a voucher had 0.07 standard deviation higher math scores after one year, 0.05 standard deviation higher math scores after three years, and 0.15 standard deviation higher math scores after four or more years. There is a less clear indication that these effects improve over time, at least when comparing the results from years one through three.

[Figure 13 about here]

The TOT math results in Figure 14 show null effects in the first year, a negative effect in the second year, and positive effects after three or more years. The negative TOT effect of two years of treatment is relatively small (-0.04 standard deviations), and primarily driven by the Louisiana Scholarship Program, which had year two impacts of -0.34 standard deviations. The positive TOT effect of three years of treatment is also relatively small (0.05 standard deviations), and is primarily driven by the Milwaukee three year impact, which was large but not statistically significant on its own. The TOT effect of four or more years of treatment, however was large (0.33 standard deviations), and relatively precise [95% CI: 0.28, 0.37]). This longer-term

outcome is primarily driven by large effects of the PACES program in Bogota, Colombia (0.80 standard deviations).

[Figure 14 about here]

In summary, these results indicate positive effects of school vouchers that vary by subject (math or reading), location (US v. non-US), and funding type (public or private). Generally, the impacts of private school vouchers are larger for reading than for math. Impacts tend to be larger for programs outside the US relative to those within the US. Impacts also generally are larger for publicly-funded programs relative to privately-funded programs. In the next section, we summarize our conclusions and explain the implications of these results in more detail.

5-D. Robustness of the Results

The effect size estimates for Bogota, Columbia seem to be an outlier. Hence, the meta-analysis was repeated after removing the data for Bogota, Columbia. The resulting overall estimates shrink in general and the conclusions are robust for both international and global ITT and TOT reading estimates. There is a significant reduction is overall ITT math estimates for international studies (-0.00 standard deviations with [95% CI: -0.05, 0.05]) however, the overall global impact is still positive and statistically significant (0.04 standard deviations with [95% CI: 0.01, 0.07]). The TOT math estimates have on overall null effect for international studies as well as a null global estimate.

The overall ITT reading effect size for publicly funded programs with Bogota, Columbia removed is 0.06 standard deviations [95% CI: -0.02, 0.15]. Hence the effect of ITT reading for

publicly funded programs was mainly driven by the Bogota, Colombia outlier. The overall TOT reading estimate for publicly funded programs has a null effect. The overall ITT math effect size for publicly funded programs with Bogota, Columbia removed is 0.12 standard deviations [95% CI: 0.03, 0.20]. Hence, the result for ITT math are robust to the removal of Bogota, Columbia outlier. The TOT math estimates have an overall negative effect (-0.15 standard deviations with [95% CI: -0.23, -0.08].

For the analysis based on years of treatment, the ITT reading estimates slightly shrink for three years of treatment. However, for four and more years of treatment, the overall estimate is statistically significant and positive and thus robust to the removal of the outlier. A similar result is obtained for TOT reading estimates. Contrary to this, the ITT math estimates are not effected for three years of treatment but have null to positive effect for four or more years of treatment. Lastly, for TOT math the overall impacts are null to positive for three years of treatment and null for four or more years of treatment.

From the robustness check, it seems that the overall conclusion for reading impacts are not affected by the Bogota, Columbia outlier but math impacts are affected negatively. The conclusions for ITT math estimates are robust to the removal of outlier for overall global estimates and publicly funded programs.

6. DISCUSSION AND CONCLUSIONS

This meta-analysis contributes to the field by combining and systematically evaluating rigorous evidence from all RCT studies of the effects of private school vouchers on student achievement. This review provides a broader overview of all the rigorous experimental findings and will have important policy implications about the effectiveness of voucher programs generally. While

voucher programs are growing across the globe, a meta-analysis of the participant effect of vouchers internationally has been lacking. As the first meta-analysis of its type, it will help establish the baseline for future studies.

We report nine meta-analytic ITT effect sizes for reading scores (six in the US and three outside of the US). For reading impacts, overall, we find positive effects of about 0.17 standard deviations (null for US programs, 0.24 standard deviations for non-US programs). A key driver of this difference is one program in Bogota, Colombia, PACES, which demonstrates very strong positive effects. Angrist et al. (2006) attempts to reconcile some of the differences between their results in Bogota and the small or null impacts in many US-based programs. It could be that there is a much larger gap in the quality of public and private schools in Colombia (and other countries, for that matter) than in the US (Angrist et al. 2006). In addition, the PACES program was distinctive in providing individual student incentives for academic achievement.

We also report 10 meta-analytic TOT effect sizes for reading (seven in the US and three outside of the US). Again, we find null effects in the US and large positive effects (0.27 standard deviations) outside of the US, primarily driven by PACES.

For math scores, we report 10 meta-analytic ITT effect sizes (seven in the US and three outside of the US). Overall, vouchers have a positive effect on math of 0.11 standard deviations, 0.07 standard deviations in the US and 0.15 standard deviations outside of the US. The TOT effects include one additional program, the Louisiana Scholarship Program (Abdulkadiroglu et al., 2015; Mills et al., 2016). The TOT math effects are a bit larger than the ITT effects overall (0.15 standard deviations). With the inclusion of Louisiana, the overall TOT effects for US programs is null, but the overall TOT effects for the non-US programs is higher at 0.33 standard deviations.

The overall results just described in this section are for the final year of data in each study. It could be that these effects are not representative of the initial effects one might expect from a new program. In fact, our analysis of the effects by year indicates that the effects of private school voucher programs often start out null in the first one or two years and then turn positive. Longer-term achievement effects, of course, are much more salient than immediate achievement effects whenever longer-term effects are available.

While the results of this meta-analysis indicate that voucher programs globally tend to positively impact test scores, perhaps particularly in countries where there is more of a private-public gap in school quality, more RCTs are needed as more voucher programs launch and operate around the globe. We especially urge more experimental evaluators to consider the impacts of vouchers on key non-cognitive outcomes such as educational attainment and civic values (e.g. Wolf et al. 2013; Wolf 2007; Angrist et al. 2006). We hope that our study will motivate researchers to do more experimental evaluations of the comprehensive effects of school vouchers to address the K-12 education gap especially in third-world countries.

We draw a few tentative policy recommendations from our study. We found that publicly-funded voucher programs show larger and clearer positive effects than privately-funded programs. Public funding could be a proxy for voucher amount, as publicly-funded vouchers tend to be of significantly greater value than privately-funded ones. Because most publicly-funded vouchers must be accepted as the full cost of educating the child, families are relieved of an additional financial burden and might therefore be more likely to keep their child enrolled in a private school long enough to realize the larger academic benefits that emerge after three or more years of private schooling. Higher-value vouchers also likely motivate a higher-quality population of private schools to participate in a voucher program. Finally, it is possible that a

higher level of quality-focused regulation of private schools exists in publicly-funded versus privately-funded voucher programs. Still, the relationship between levels of regulation and the achievement benefits of vouchers remains an important but understudied question.

Additionally, in terms of policy implications, it is critical to consider the cost-benefit tradeoffs associated with voucher programs. Wolf & McShane (2013) and Muralidharan et al. (2015) found that vouchers are cost effective, since they tend to generate achievement outcomes that are as good as or better than traditional public schools but at a fraction of the cost. The greater efficiency of school choice in general and school vouchers in particular are another fruitful avenue for scholarly inquiry.

REFERENCES

Abdulkadiroglu, A., Pathak, P. A., & Walters, C. R. (2015). *School vouchers and student achievement: First-year evidence from the Louisiana Scholarship Program* (NBER Working Paper No. 21839). Cambridge, MA: National Bureau of Economic Research.

Anderson, M. R., Guzman, T., & Ringquist, E. J. (2013). Evaluating the effectiveness of educational vouchers. In E. J. Ringquist (Ed.), *Meta-analysis for public management and policy* (pp. 311-351), San Francisco, CA: Jossey-Bass.

Angrist, J., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2002). Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. *American Economic Review*, *92*(5), 1535-1558.

Angrist, J., Bettinger, E., & Kremer, M. (2006). Long-term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia. *The American Economic Review*, *96*(3), 847-862.

Barnard, J., Frangakis, C. E., Hill, J. L., & Rubin, D. B. (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *Journal of the American Statistical Association*, *98(462),* 299-323.

Bettinger, E., & Slonim, R. (2003). *The effect of educational vouchers on academic and non-academic outcomes: Using experimental economic methods to study a randomized natural experiment*. Mimeo, Case Western Reserve University.

Bettinger, E., & Slonim, R. (2006). Using experimental economics to measure the effects of a natural educational experiment on altruism. *Journal of Public Economics*, *90*(8), 1625-1648.

Betts, J.R., Tang, Y. E., & Zau, A. C. (2010). Madness in the method? A critical analysis of popular methods of estimating the effect of charter schools on student achievement. *Taking Measure of Charter Schools: Better Assessments, Better Policy Making, Better Schools,* Lanham, MD: Rowman & Littlefield Publishers, Inc.

Bitler, M., Domina, T., Penner, E. K., & Hoynes, H. (2015). Distributional effects of a school voucher program: Evidence from New York City. *Journal of Research on Education Effectiveness, 8*(3), 419-450.

Borenstein, M., Hedges, L., & Rothstein, H. (2007). *Meta-analysis fixed effect vs. random effects*. Retrieved from: https://www.meta-analysis.com/downloads/Meta-analysis%20fixed%20effect%20vs%20random%20effects.pdf

Cowen, J. M. (2008). School choice as a latent variable: Estimating the "complier average causal effect" of vouchers in Charlotte. *Policy Studies Journal, 36*(2), 301-315.

Coulson, A. J. (2009). Comparing public, private, and market schools: The international evidence. *Journal of School Choice, 3*(1), 31-54.

Dixon, P. (2013). *International aid and private schools for the poor: Smiles, miracles and markets*, Northampton, MA: Edward Elgar.

Epple, D., Romano, R. E., & Urquiola, M. (2015). *School vouchers: A survey of the economics literature* (NBER Working Paper No. 21523). Cambridge, MA: National Bureau of Economic Research.

Forster, G. (2011). *A win-win solution: The empirical evidence on school vouchers*. Indianapolis: Friedman Foundation for Educational Choice.

Forster, G. (2013). *A win-win solution: The empirical evidence on school choice*. Indianapolis: Friedman Foundation for Educational Choice.

Frendewey, M., Sawatka, K., Marcavage, W., Carney, K., Martinez, K, & Dauphin, P. (2015). *School choice yearbook 2014-2015: Breaking down barriers to choice*. Alliance for School Choice.

Friedman, M. (1955). The role of government in education. In R. A. Solo (Ed.), *Economics and the public interest* (pp. 123–144). New Brunswick, NJ: Rutgers University Press.

Glenn, C. L. (1989). *Choice of schools in six nations: France, Netherlands, Belgium, Britain, Canada, West Germany*. University of Michigan.

Glenn, C. L., De Groof, J., & Candal, C. S. (2012a). *Balancing freedom, autonomy and accountability in education*. Volume 1. Wolf Legal Publishers.

Glenn, C. L., De Groof, J., & Candal, C. S. (2012b). *Balancing freedom, autonomy and accountability in education*. Volume 2. Wolf Legal Publishers.

Glenn, C. L., De Groof, J., & Candal, C. S. (2012c). *Balancing freedom, autonomy and accountability in education*. Volume 3. Wolf Legal Publishers.

Glenn, C. L., De Groof, J., & Candal, C. S. (2012d). *Balancing freedom, autonomy and accountability in education*. Volume 4. Wolf Legal Publishers.

Greene, J. P. (2000). The effect of school choice: An evaluation of the Charlotte children's scholarship fund program. *Manhattan Institute for Policy Research Civic Report,* (12).

Greene, J. P., Peterson, P. E., & Du, J. (1999). Effectiveness of school choice: The Milwaukee experiment. *Education and Urban Society, 31(2),* 190-213.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Howell, W. G., & Peterson, P. E. (2006). *The education gap: Vouchers and urban schools*. Washington, DC: Brookings Institution Press.

Howell, W. G., Wolf, P. J., Campbell, D. E., & Peterson, P. E. (2002). School vouchers and academic performance: Results from three randomized field trials. *Journal of Policy Analysis and Management, 21(2),* 191-217.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Jin, H., Barnard, J., & Rubin, D. B. (2010). A modified general location model for noncompliance with missing data revisiting the New York City school choice scholarship program using principal stratification. *Journal of Educational and Behavioral Statistics, 35(2),* 154-173.

Krueger, A. B., & Zhu, P. (2004). Another look at the New York City school voucher experiment. *American Behavioral Scientist, 47*(5), 658-698.

Lubienski, C., & Weitzel, P. (2008). The effects of vouchers and private schools in improving academic achievement: A critique of advocacy research. *Brigham Young University Law Review, 2008*(2), 447-485.

Mayer, D. P., Peterson, P. E., Myers, D. E., Tuttle, C. C., & Howell, W. G. (2002). *School choice in New York City after three years: An evaluation of the school choice scholarships program.* Washington, DC: Mathematica Policy Research, Inc.

Mill, J. S. (1962 [1869]) *Utilitarianism, on liberty, essay on Bentham*. (Warnock, M. ed.) Meridian.

Mills, J. N., & Wolf, P. J. (2016). The effects of the Louisiana scholarship program on student achievement after two years. *Available at SSRN 2738805*.

Miron, G., Evergreen, S., Urschel, J. (2008). *The impact of school choice reforms on student achievement*. The Great Lakes Center for Education Research & Practice.

Mizala, A., & Romaguera, P. (2000). School performance and choice: The Chilean experience. *The Journal of Human Resources, 35*(2), 392-417.

Mosteller, F. & Boruch, F. (Eds.) (2002). *Evidence matters: Randomized trials in education research.* Washington, D.C.: The Brookings Institution.

Muralidharan, K., & Sundararaman. V. (2015). The aggregate effect of school choice evidence from a two-stage experiment in India. *The Quarterly Journal of Economics, 130*(3), 1011-1066.

Paine, T. (1791). *The rights of man: Answer to Mr. Burke's attack on the French revolution*. J. S. Jordan.

Peterson, P. E., Howell, W. G., Wolf, P. J., & Campbell, D. E. (2003). School vouchers: Results from randomized experiments. In Caroline M. Hoxby (Ed.), *The Economics of School Choice* (pp. 107-144). University of Chicago Press.

PROBE Team. (1999). *Public report on basic education in India*. Oxford, Eng.: Oxford University Press.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach*. Seventh edition. Sage.

Rouse, C. E. (1998). Private school vouchers and student achievement: An evaluation of the Milwaukee parental choice program. *Quarterly Journal of Economics, 113(2),* 553-602.

Rouse, C. E., & Barrow, L. (2008). *School vouchers and student achievement: Recent evidence and remaining questions*. National Center for the Study of the Privatization of Education.

Rouse, C. E., & Barrow, L. (2009). School vouchers and student achievement: Recent evidence and remaining questions. *Annual Review of Economics, 1*(1), 17–42.

Tooley, J. (2009). *The beautiful tree: A personal journey into how the world's poorest people are educating themselves*, Washington, D. C.: Cato Institute.

Usher, A., & Kober, N. (2011). *Keeping informed about school vouchers: A review of major developments and research.* Center on Education Policy.

Wolf, P. J. (2007). Civics exam: Schools of choice boost civic values. *Education Next, 7*(3), 66-72.

Wolf, P. J. (2008a). Vouchers. In *The International Encyclopedia of Education*, (pp. 635-36). Routledge.

Wolf, P. J. (2008b). School voucher programs: What the research says about parental school choice. *Brigham Young University Law Review*, *2008*(2), 415-446.

Wolf, P. J., Egalite, A. J., & Dixon, P. (2015). Private school choice in developing countries: Experimental results from Delhi, India. *Handbook of International Development and Education*, 456-471.

Wolf, P. J., Kisida, B., Gutmann, B., Puma, M., Eissa, N., & Rizzo, L. (2013). School vouchers and student outcomes: Experimental evidence from Washington, DC. *Journal of Policy Analysis and Management*, *32(2)*, 246-270.

Wolf, P. J., & Macedo, S., eds. (2004). *Educating citizens: International perspectives on civic values and school choice*. Brookings.

Wolf, P. J. & McShane, M. (2013). Is the juice worth the squeeze? A benefit/cost analysis of the District of Columbia opportunity scholarship program. *Education Finance and Policy, 8*(1), 74–99.

TABLES AND FIGURES

Table 1: U.S. Empirical Studies of School Vouchers Included in or Ignored by Literature Reviewers

Study	Program	Nature of Achievement Findings	Miron, Evergreen & Urschel 2008	Lubienski & Weitzel 2008	Wolf 2008	Rouse & Barrow 2008	Coulson 2009	CEP 2011	Forster 2011	Forster 2013	Anderson, Guzman & Ringquist 2013	Epple, Romano & Urquiola 2015	Total	Percent
Witte 1995; 1998; 2000	Milwaukee	No impacts	x	x			x				x	x	5	50%
Rouse 1998	Milwaukee	**Pos in math**	x	x	x	x	x		x	x	x	x	9	90%
Greene, Peterson & Du 1996; 1999	Milwaukee	**Pos in reading & math**	x	x	x		x		x	x	x		7	70%
Greene, Howell & Peterson 1998	Cleveland	Pos for subgroups in reading, math, science	x	x							x		3	30%
Greene 2000	Charlotte	**Pos in reading & math**			x				x	x	x		4	40%
Metcalf 2003	Cleveland	No impacts	x	x			x	x			x		5	50%
Peterson et al 2003; Howell et al. 2002	Dayton	**Pos for African Americans**	x	x	x	x	x		x	x	x	x	9	90%
Peterson et al 2003; Howell et al. 2002	DC	**No impacts**	x	x	x	x	x		x	x	x	x	9	90%
Peterson et al 2003; Mayer et al. 2002	NYC	**Pos for African Americans**	x	x	x	x	x		x	x	x	x	9	90%
Barnard et al 2003	NYC	**Pos in math for African Americans**			x				x	x	x	x	5	50%
Kruger & Zhu 2004	NYC	**No impacts**	x	x	x	x	x		x	x	x	x	9	90%
Bettinger & Slonim 2006	Toledo	**No impacts**											0	0%
Plucker et al 2006	Cleveland	No impacts	x					x			x		3	30%
Belfield 2006	Cleveland	Neg in math	x	x		x	x	x			x		6	60%
Wolf et al 2007	DC	**No impacts**	x		x	x	x	x	x	x			7	70%
Cowen 2008	Charlotte	**Pos in reading & math**			x				x	x	x		4	40%
Wolf et al 2008	DC	**No impacts**				x	x	x	x	x	x		6	60%
Wolf et al 2009	DC	**Pos in reading**						x	x	x	x		4	40%
Witte et al 2009	Milwaukee	No impacts						x				x	2	20%
Wolf et al 2010; 2013	DC	**Pos for subgroups in reading**						x	x	x	x	x	5	50%
Jin, Barnard & Rubin 2010	NYC	**Pos in math**								x			1	10%
Witte et al 2010	Milwaukee	No impacts						x				x	2	20%
Witte et al 2011	Milwaukee	No impacts						x				x	2	20%
Figlio 2011	Florida	Pos in reading											0	0%
Witte et al 2012	Milwaukee	Pos in reading										x	1	10%
Bitler et al. 2015	NYC	**No impacts**											0	0%
Abdulkadiroglu et al 2015	Louisiana	**Neg in math**											0	0%
Mills & Wolf 2016	Louisiana	**Neg in math**											0	0%
Total			**12**	**10**	**10**	**8**	**11**	**10**	**13**	**14**	**17**	**12**	**117**	
Percent of Targets			**80.0%**	**62.5%**	**90.9%**	**61.5%**	**64.7%**	**52.6%**	**86.7%**	**93.3%**	**68.0%**	**48.0%**		

Notes: Dark borders of the columns demark the time period of the author search. Shaded cells indicate studies that were excluded due to scientifically valid exclusion criteria. Items in bold in the "Findings" column signify results from RCTs.

Table 2: Description of 19 RCT Studies included in Meta-Analysis

Authors	Publication Year	Years of Treatment	Program Evaluated	Duration of Study	Grades	(First Outcome Year)	Program Attrition (Final Year)	Sample Attrition (Final Year)
Abdulkadiroglu, Pathak & Walters	2015	1	Louisiana Scholarship Program (LSP)	2012-2013 (1 year)	3 to 8	N/A	N/A	N/A
Angrist, Bettinger, Bloom, King & Kremer	2002	3	Programa de Ampliacion de Cobertura de la Educacion Secundaria (PACES)	1995-1999 (4 years)	6 to 9	283	10%	75.3%
Angrist, Bettinger, & Kremer	2006	7	Programa de Ampliacion de Cobertura de la Educacion Secundaria (PACES)	1994-2001 (8 years)	6 to 11	3,541	50%	12.4%
Barnard, Frangakis, Hill & Rubin	2003	1	The School Choice Scholarships Foundation Program	1997-2000 (4 years)	1 to 4	525	23.5%	22.3%
Bettinger & Slonim	2006	3	Children's Scholarship Fund	1998-2001 (4 years)	K to 8	186	N/A	92%
Bitler, Domina, Penner & Hoynes	2015	3	New York City School Choice Program	1997-2000 (4 years)	K to 4	2,080	41.3%	34.6% Reading; 35.0% Math
Cowen	2008	1	Charlotte Children's Scholarship Fund	1999-2000 (1 year)	2 to 8	347	25.5%	70%
Greene	2000	1	Charlotte Children's Scholarship Fund	1999-2000 (1 year)	2 to 8	357	51.6%	60%
Greene, Peterson & Du	1999	4	Milwaukee Parental Choice Program (MPCP)	1990-1994 (5 years)	K to 8	816	N/A	60% Treatment, 52% Control
Howell, Wolf, Campbell & Peterson	2002	3	The School Choice Scholarships Foundation Program	1997-2000 (4 years)	1 to 4	1,434	N/A	33%
Howell, Wolf, Campbell & Peterson	2002	2	Parents Advancing Choice in Education	1998-2000 (2 years)	K to 12	404	N/A	51%
Howell, Wolf, Campbell & Peterson	2002	3	Washington Scholarship Fund	1998-2001 (3 years)	K to 8	930	76%	40%
Jin, Barnard & Rubin	2010	1	New York City School Choice Program	1997-2000 (4 years)	1 to 4	525	23.5%	22.3%
Krueger & Zhu	2004	3	New York City School Choice Program	1997-2000 (4 years)	K to 4	2,080	41.3%	36.2%
Mills & Wolf	2015	2	Louisiana Scholarship Program (LSP)	2012-2014 (2 years)	3 to 8	N/A	N/A	N/A
Muralidharan & Sundararaman	2015	4	Andhra Pradesh (AP) School Choice Experiment	2008-2012 (4 years)	1 to 5	4,620	49%	20.7% English; 68.1% Hindi; 17.5% Telugu; 17.5% Math
Rouse	1998	4	Milwaukee Parental Choice Program (MPCP)	1990-1994 (5 years)	K to 8	1,343	75.5%	N/A
Wolf, Egalite & Dixon	2012	2	Ensure Access to Better Learning Experiences (ENABLE)	2011-2013 (2 years)	K to 2	1,306	11%	N/A
Wolf, Kisida, Gutmann, Puma, Eissa & Rizzo	2013	4	District of Columbia OpportunityScholarship Program (OSP)	2004-2009 (6 years)	K to 12	1,649	17.9%	37.8% Treatment, 48.5% Control

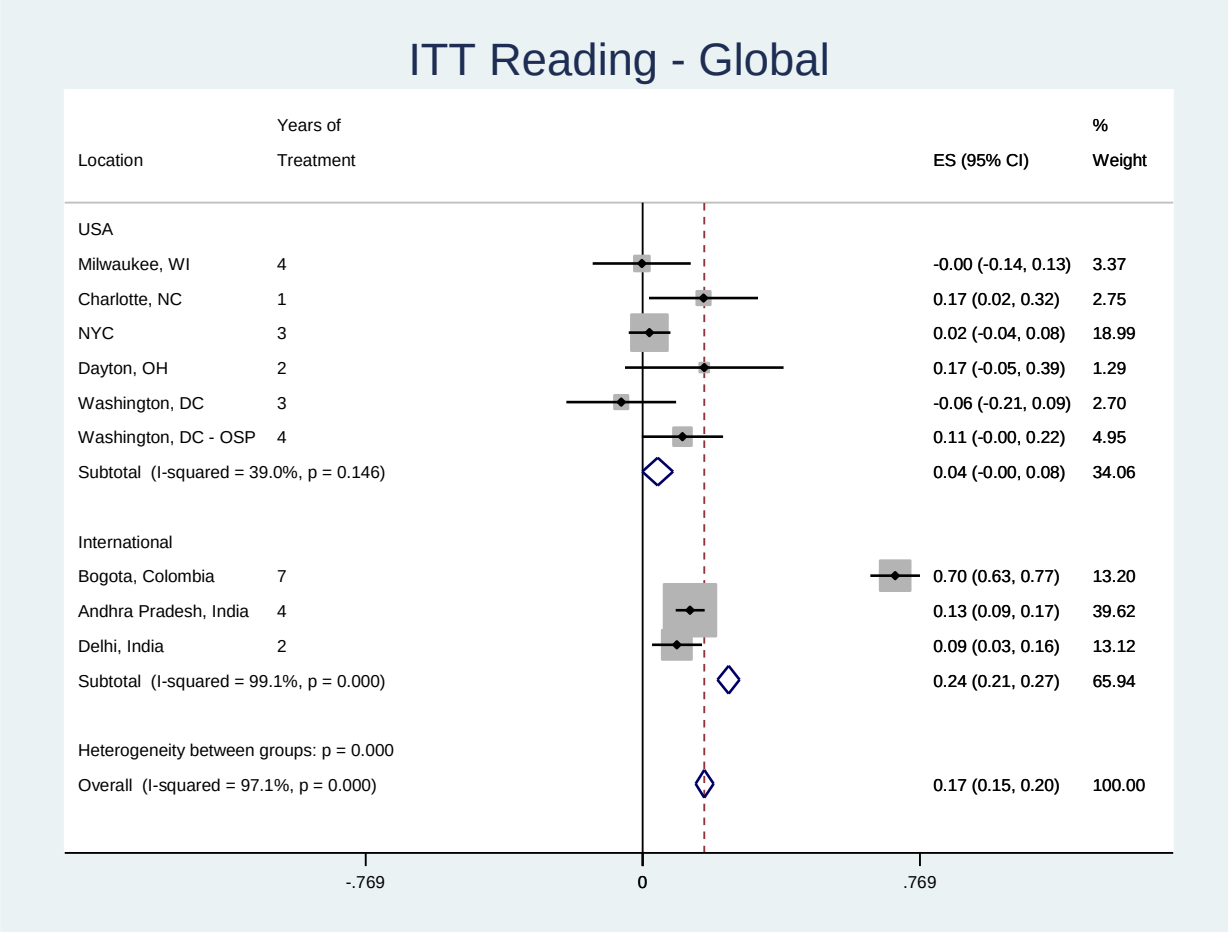
Notes: The sample size and attrition rates are based on the estimates from ITT Reading with the exception of Bettinger & Slonim (2006) which had only math impacts. The actual sample sizes for calculating the ITT and TOT Reading and Math impacts may differ slightly.

Table 3: Description of 11 Voucher Programs included in Meta-Analysis

Program Evaluated	Location	Funding Source	Funding Amount (Full or Partial)	Grades	Studies Cited
Andhra Pradesh (AP) School Choice Experiment	Andhra Pradesh, India	Private	Full	1 to 5	Muralidharan & Sundararaman (2015)
Charlotte Children's Scholarship Fund	Charlotte, NC (USA)	Private	Partial	2 to 8	Greene (2000); Cowen (2008)
Children's Scholarship Fund	Toledo, OH (USA)	Private	Partial	K to 8	Bettinger & Slonim (2006)
District of Columbia Opportunity Scholarship Program (OSP)	Washington, DC (USA)	Public	Full	K to 12	Wolf, Kisida, Gutmann, Puma, Eissa & Rizzo (2013)
Ensure Access to Better Learning Experiences (ENABLE)	Delhi, India	Private	Full	K to 2	Wolf, Egalite & Dixon (2015)
Louisiana Scholarship Program (LSP)	Louisiana (USA)	Public	Full	3 to 8	Abdulkadiroglu, Pathak & Walters (2015); Mills & Wolf (2016)
Milwaukee Parental Choice Program (MPCP)	Milwaukee, WI (USA)	Public	Full	K to 8	Rouse (1998); Greene, Peterson & Du (1999)
Parents Advancing Choice in Education	Dayton, OH (USA)	Private	Partial	K to 12	Peterson, Howell, Wolf & Campbell (2003)
Programa de Ampliacion de Cobertura de la Educacion Secundaria (PACES)	Bogota, Colombia	Public (partly funded by World Bank)	Partial	6-9 (2002 paper) and 6-11 (2006 paper)	Angrist, Bettinger, Bloom, King & Kremer (2002); Angrist, Bettinger, & Kremer (2006)
School Choice Scholarships Foundation	New York, NY (USA)	Private	Partial	1 to 4	Peterson, Howell, Wolf & Campbell (2003); Barnard, Frangakis, Hill & Rubin (2003); Krueger & Zhu (2004); Jin, Barnard & Rubin (2010); Bitler, Domina, Penner & Hoynes (2015)
Washington Scholarship Fund	Washington, DC (USA)	Private	Partial	K to 8	Peterson, Howell, Wolf & Campbell (2003)

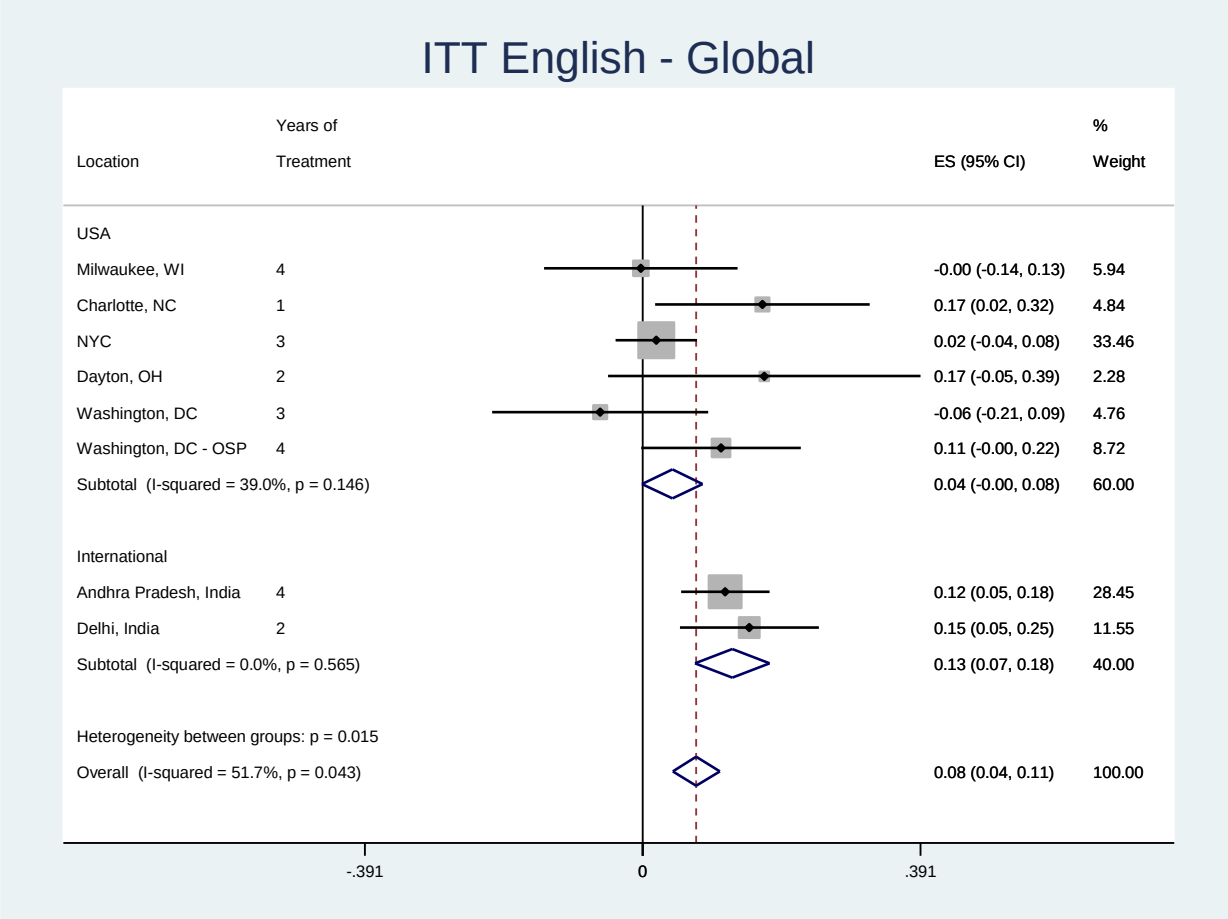
Note: Studies do not necessarily contain all years of a program. See Table 2 for more details at the study level.

Figure 1:



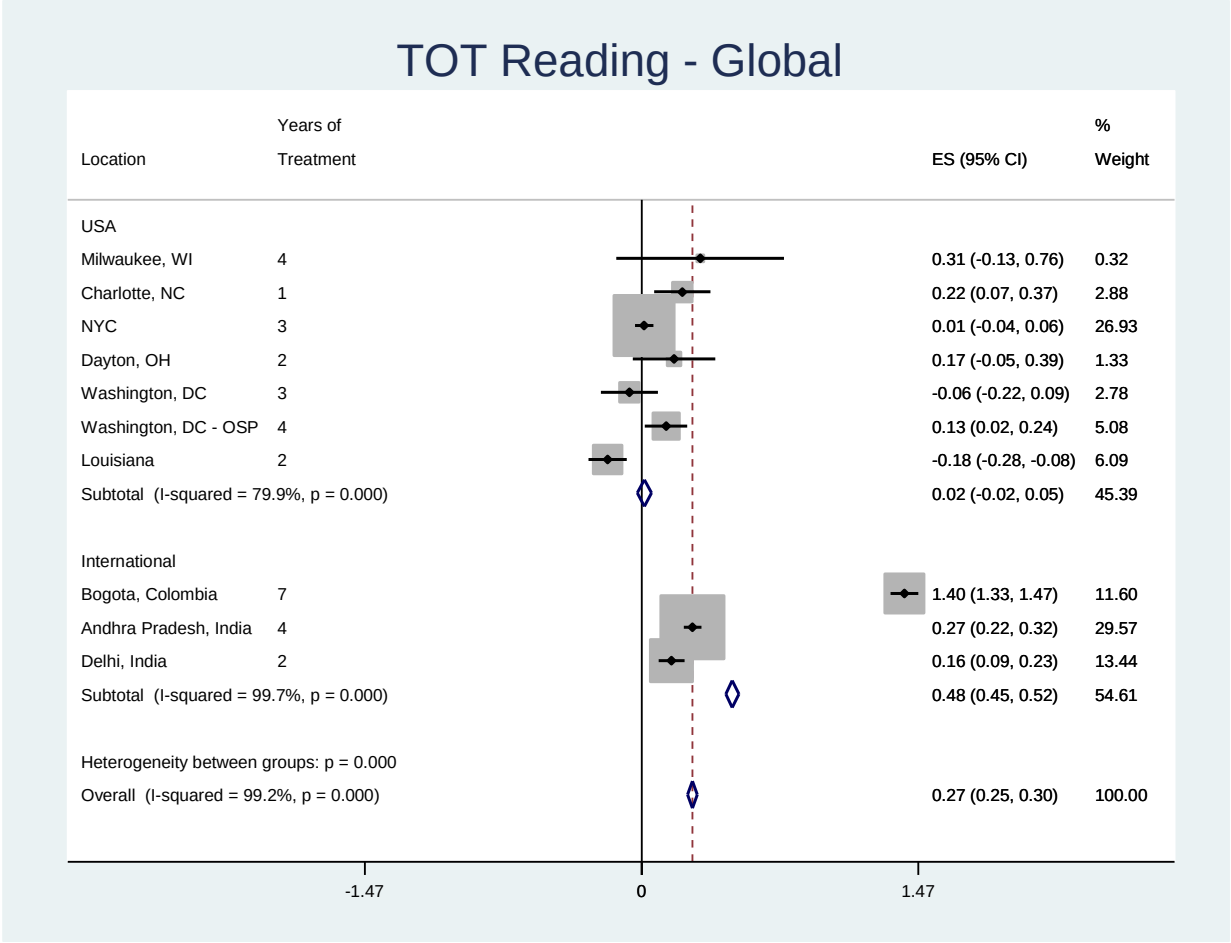
Note: The Hedges' *g* estimates are based on last year effect size calculated for each study. The boxes show overall estimates for USA, International (outside USA) and Global (red dotted line). The grey area around each point (effect size) is the weight of each study (inverse of variance). No reading estimates are reported for Toledo, OH as it had only math test outcomes. Reading estimate for Delhi, India includes an overall estimate for English and Hindi. Reading estimate for Andhra Pradesh, India includes an overall estimate for English, Hindi and Telugu. Reading estimate for Bogota, Colombia is for Spanish. Louisiana voucher program did not have ITT estimates as it was a placement lottery. Overall effect size for International studies with Bogota, Columbia removed is 0.12 (0.09, 0.16) and overall global average is 0.09 (0.06, 0.12).

Figure 2:



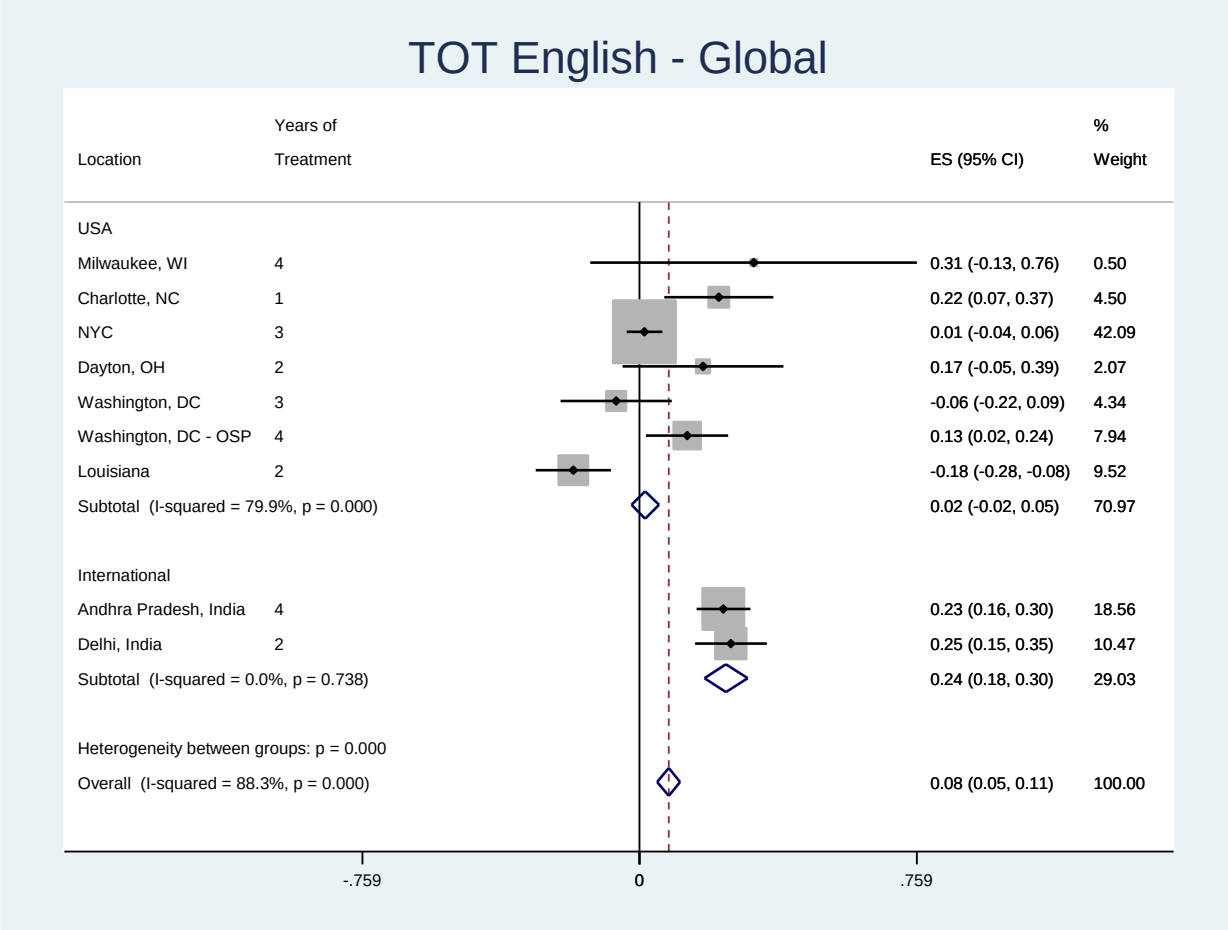
Note: The Hedges' *g* estimates are based on last year effect size calculated for each study. The boxes show overall estimates for USA, International (outside USA) and Global (red dotted line). The grey area around each point (effect size) is the weight of each study (inverse of variance). No reading estimates are reported for Toledo, OH as it had only math test outcomes. Bogota, Colombia did not have an English estimate as the tests were administered in Spanish. Louisiana voucher program did not have ITT estimates as it was a placement lottery.

Figure 3:



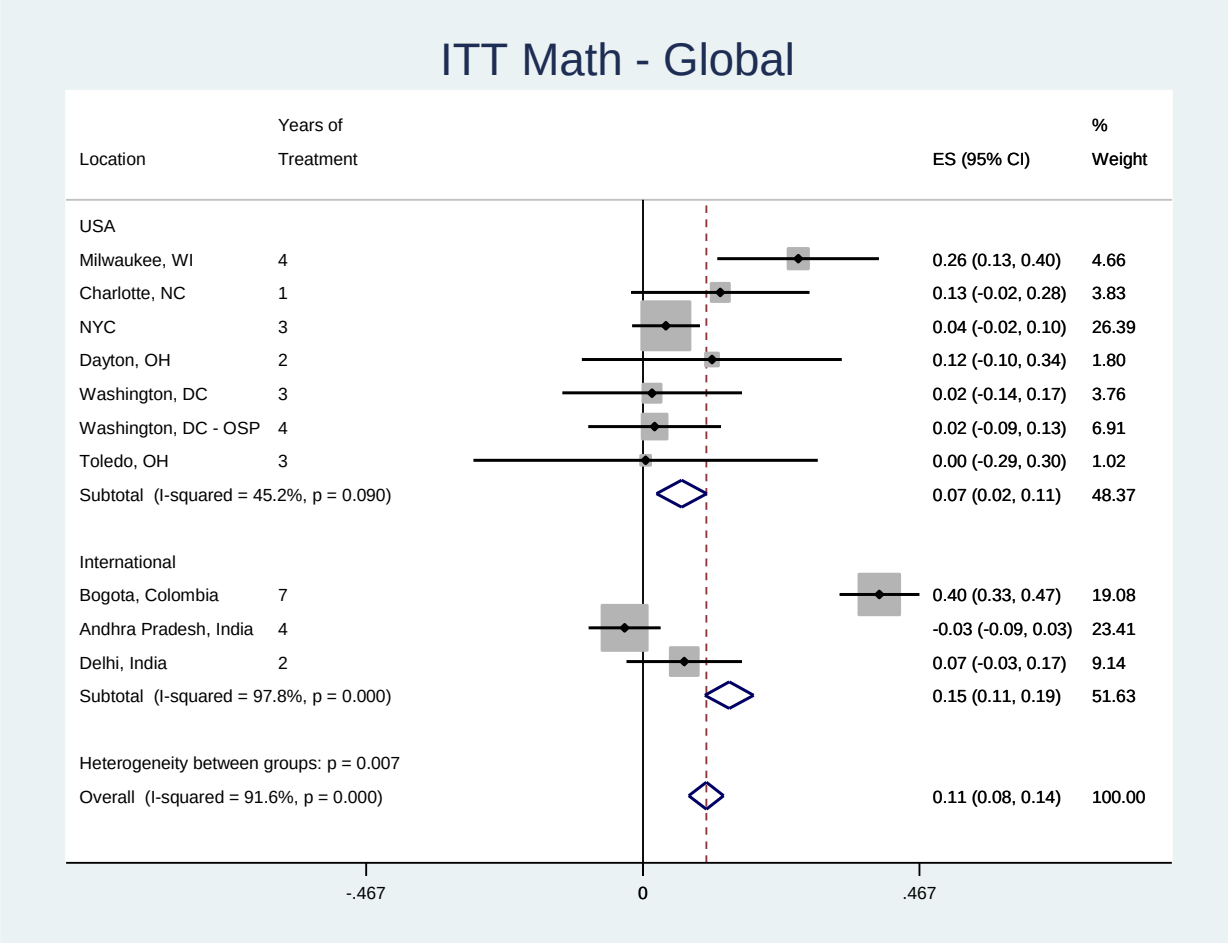
Note: The Hedges' *g* estimates are based on last year effect size calculated for each study. The boxes show overall estimates for USA, International (outside USA) and Global (red dotted line). The grey area around each point (effect size) is the weight of each study (inverse of variance). No reading estimates are reported for Toledo, OH as it had only math test outcomes. Reading estimate for Delhi, India includes an overall estimate for English and Hindi. Reading estimate for Andhra Pradesh, India includes an overall estimate for English, Hindi and Telugu. Reading estimate for Bogota, Colombia is for Spanish. Overall effect size for International studies with Bogota, Columbia removed is 0.24 (0.20, 0.27) and overall global average is 0.12 (0.10, 0.15).

Figure 4:



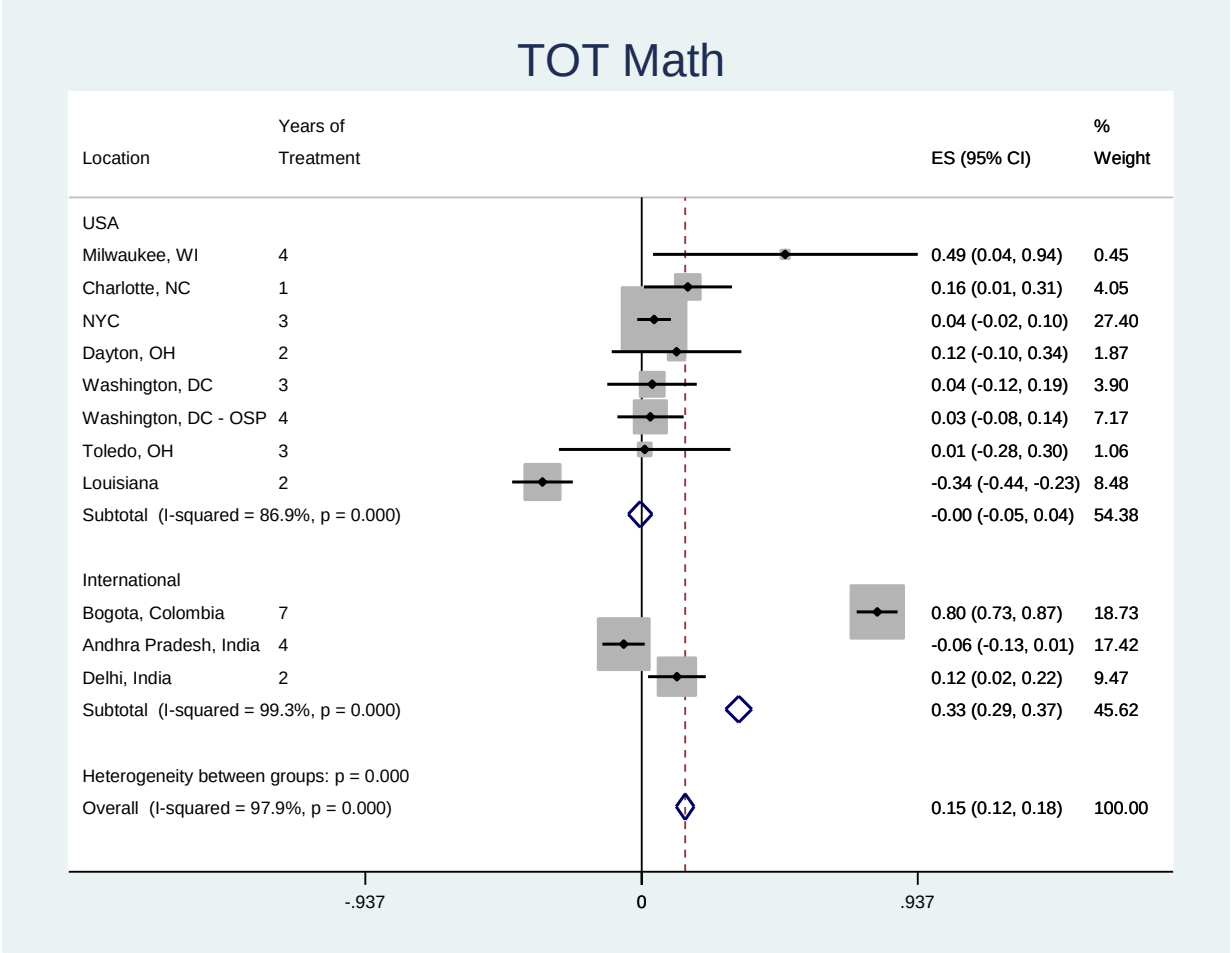
Note: The Hedges' *g* estimates are based on last year effect size calculated for each study. The boxes show overall estimates for USA, International (outside USA) and Global (red dotted line). The grey area around each point (effect size) is the weight of each study (inverse of variance). No reading estimates are reported for Toledo, OH as it had only math test outcomes. Bogota, Colombia did not have an English estimate as the tests were administered in Spanish.

Figure 5:



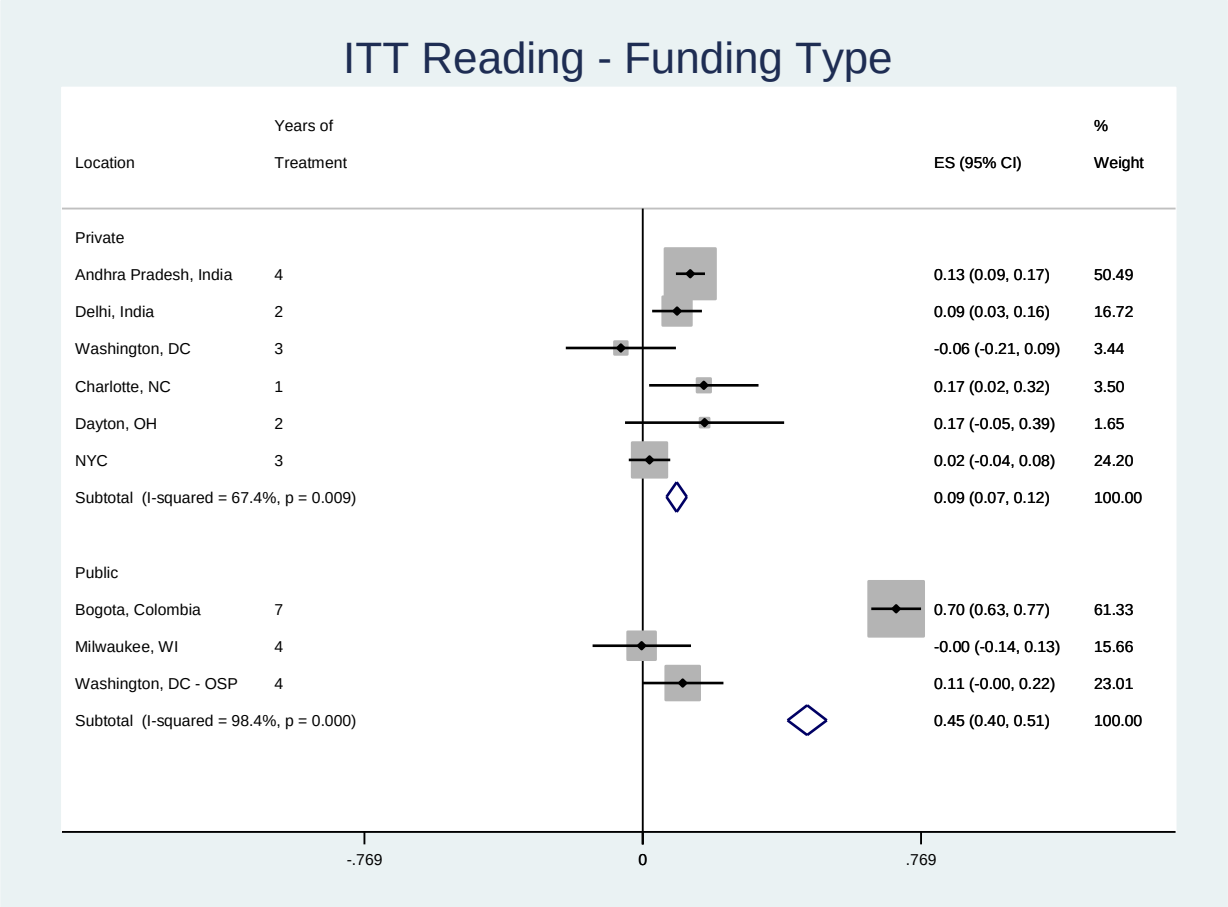
Note: The Hedges' *g* estimates are based on last year effect size calculated for each study. The boxes show overall estimates for USA, International (outside USA) and Global (red dotted line). The grey area around each point (effect size) is the weight of each study (inverse of variance). Louisiana voucher program did not have ITT estimates as it was a placement lottery. Overall effect size for International studies with Bogota, Columbia removed is -0.00 (-0.05, 0.05) and overall global average is 0.04 (0.01, 0.07).

Figure 6:



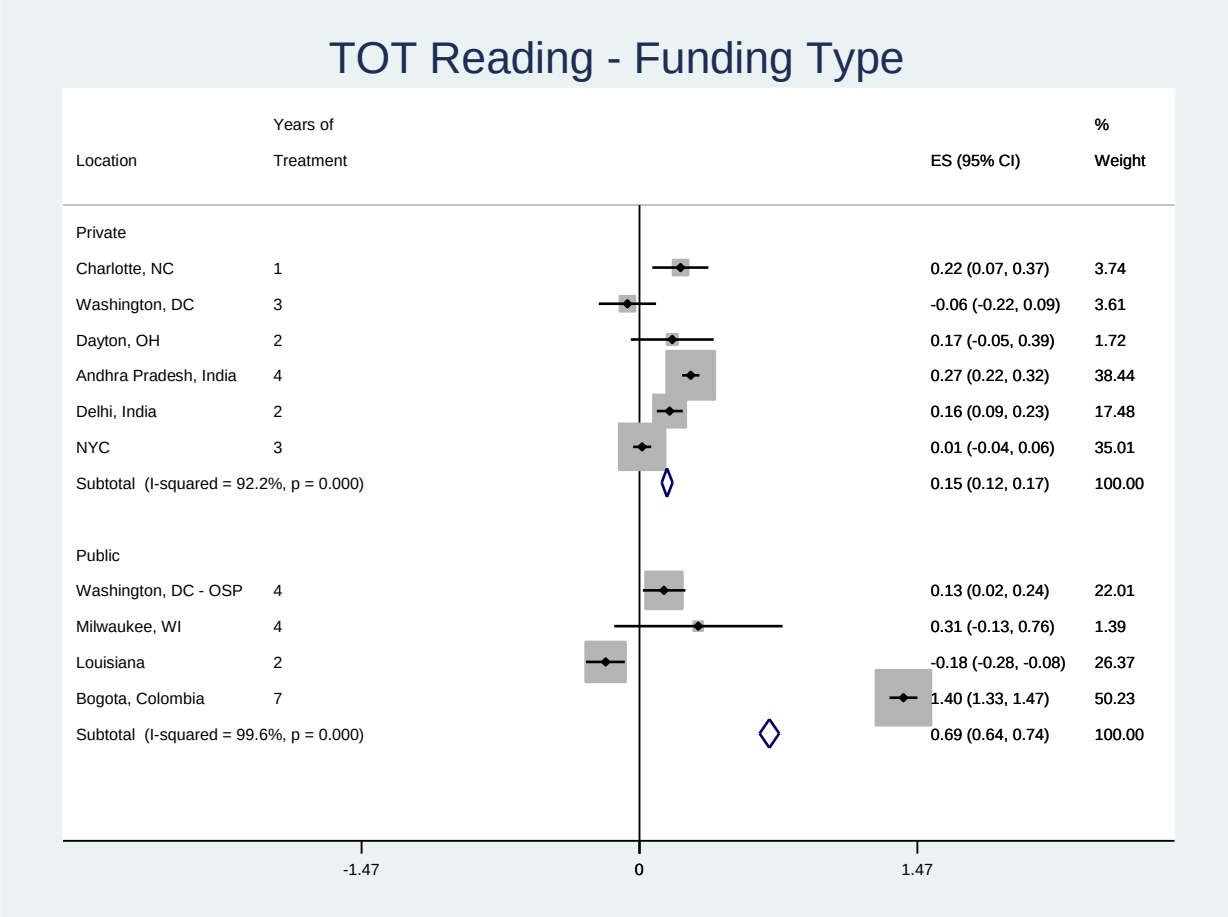
Note: The Hedges' *g* estimates are based on last year effect size calculated for each study. The boxes show overall estimates for USA, International (outside USA) and Global (red dotted line). The grey area around each point (effect size) is the weight of each study (inverse of variance). Overall effect size for International studies with Bogota, Columbia removed is 0.00 (-0.06, 0.06) and overall global average is -0.00 (-0.04, 0.03).

Figure 7:



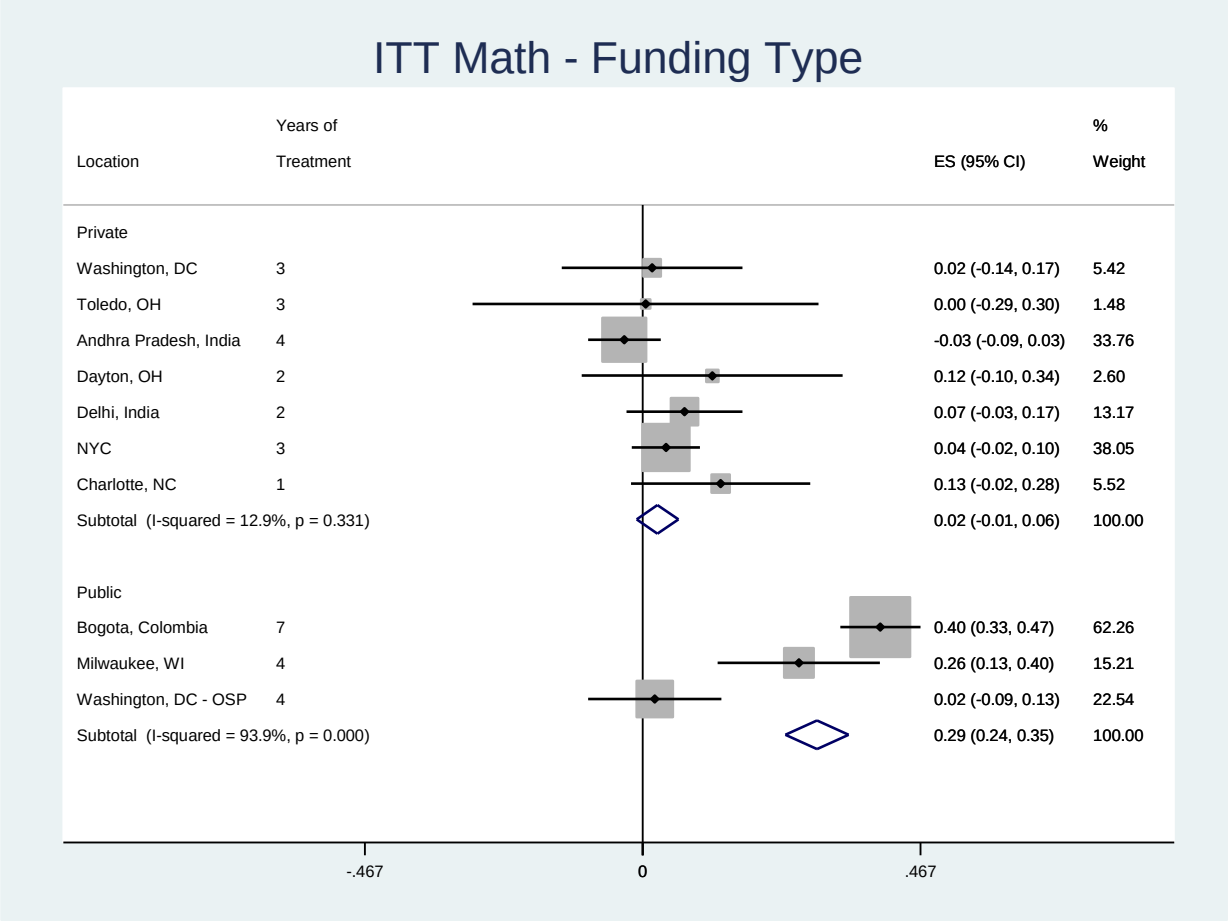
Note: The Hedges' *g* estimates are based on last year effect size calculated for each study. The boxes show overall estimates for privately and publicly (having received any public funds) funded programs. The grey area around each point (effect size) is the weight of each study (inverse of variance). No reading estimates are reported for Toledo, OH as it had only math test outcomes. Reading estimate for Delhi, India includes an overall estimate for English and Hindi. Reading estimate for Andhra Pradesh, India includes an overall estimate for English, Hindi and Telugu. Reading estimate for Bogota, Colombia is for Spanish. Louisiana voucher program did not have ITT estimates as it was a placement lottery. Overall effect size for publicly funded programs with Bogota, Columbia removed is 0.06 (-0.02, 0.15).

Figure 8:



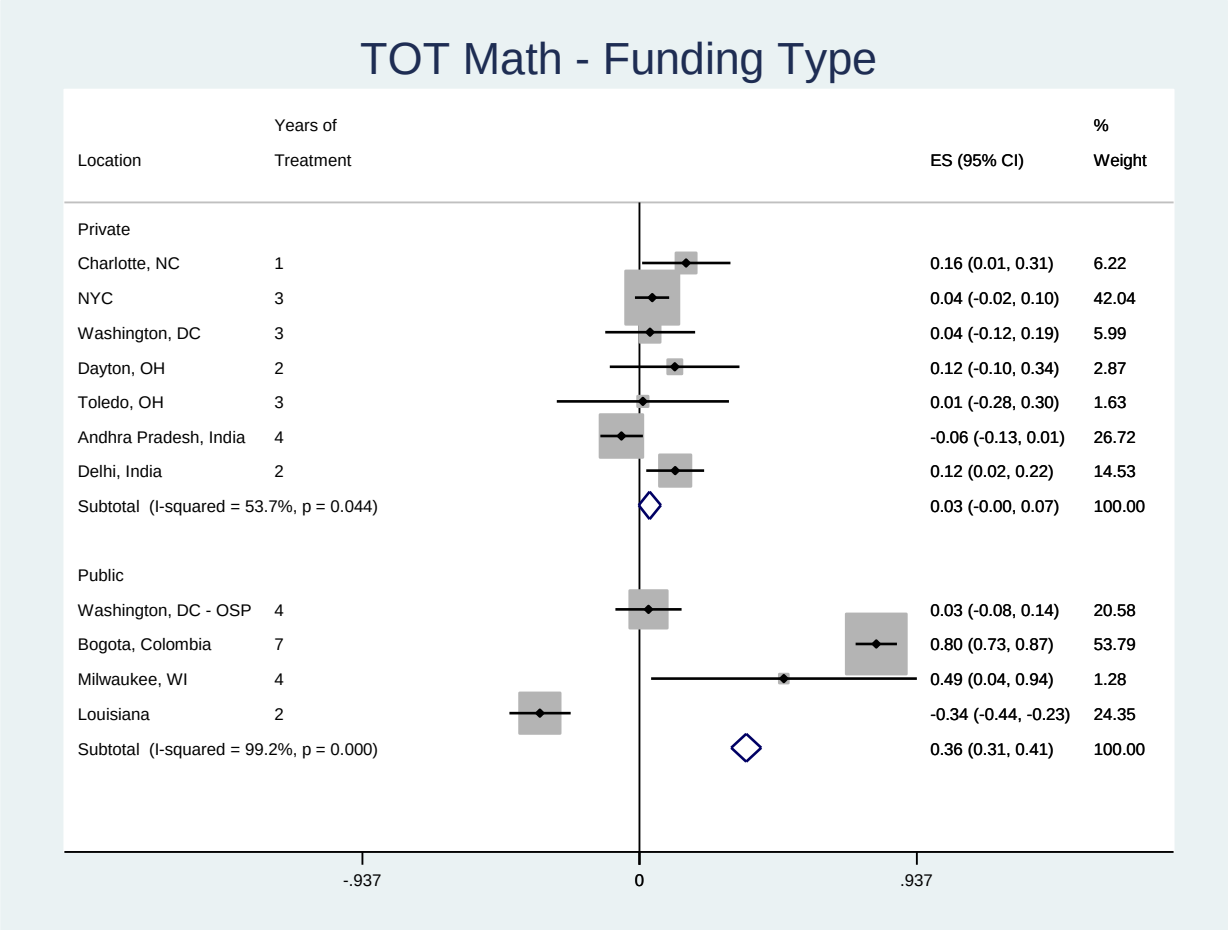
Note: The Hedges' *g* estimates are based on last year effect size calculated for each study. The boxes show overall estimates for privately and publicly (having received any public funds) funded programs. The grey area around each point (effect size) is the weight of each study (inverse of variance). No reading estimates are reported for Toledo, OH as it had only math test outcomes. Reading estimate for Delhi, India includes an overall estimate for English and Hindi. Reading estimate for Andhra Pradesh, India includes an overall estimate for English, Hindi and Telugu. Reading estimate for Bogota, Colombia is for Spanish. Overall effect size for publicly funded programs with Bogota, Columbia removed is -0.03 (-0.10, 0.04).

Figure 9:



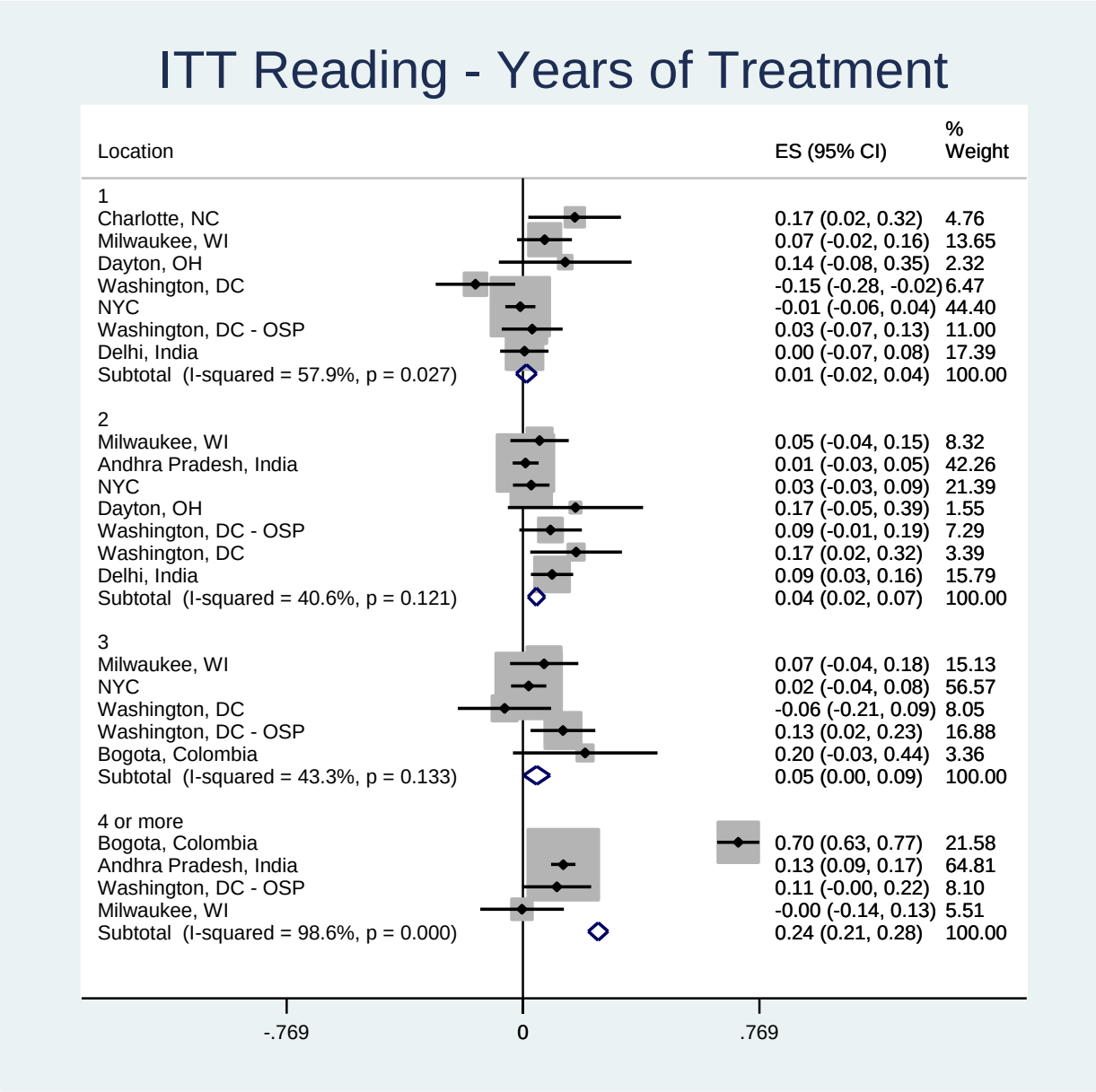
Note: The Hedges' *g* estimates are based on last year effect size calculated for each study. The boxes show overall estimates for privately and publicly (having received any public funds) funded programs. The grey area around each point (effect size) is the weight of each study (inverse of variance). Louisiana voucher program did not have ITT estimates as it was a placement lottery. Overall effect size for publicly funded programs with Bogota, Columbia removed is 0.12 (0.03, 0.20).

Figure 10:



Note: The Hedges' *g* estimates are based on last year effect size calculated for each study. The boxes show overall estimates for privately and publicly (having received any public funds) funded programs. The grey area around each point (effect size) is the weight of each study (inverse of variance). Overall effect size for publicly funded programs with Bogota, Columbia removed is -0.15 (-0.23, -0.08).

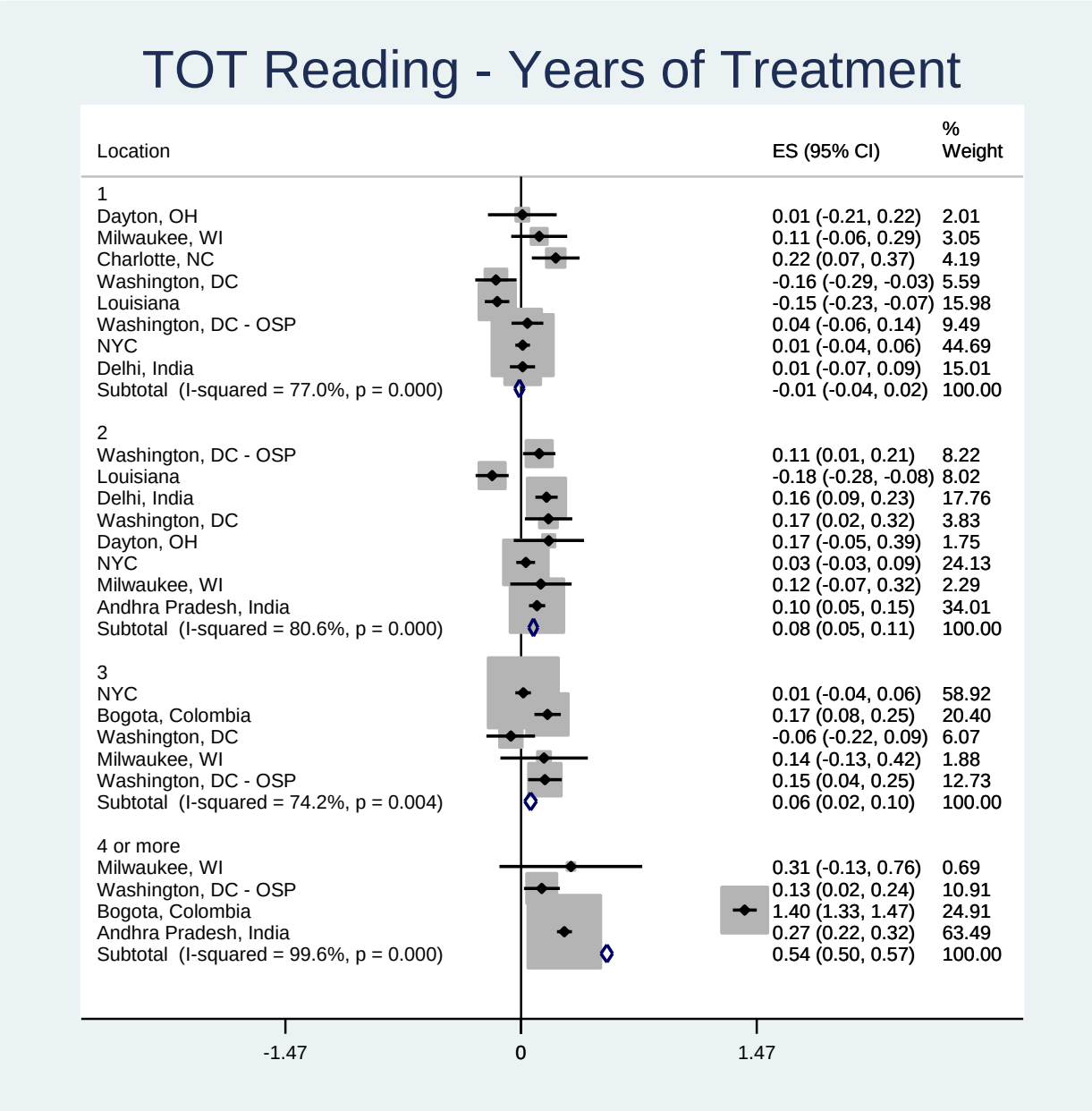
Figure 11:



Note: The Hedges' *g* estimates are based on one year effect, two year effect, three year effect and four or more year effect size calculated for each study. The boxes show overall estimates for yearly effect of programs. The grey area around each point (effect size) is the weight of each study (inverse of variance). No reading estimates are reported for Toledo, OH as it had only math test outcomes. Reading estimate for Delhi, India includes an overall estimate for English and Hindi. Reading estimate for Andhra Pradesh, India includes an overall estimate for English, Hindi and Telugu. Reading estimate for Bogota, Colombia is for Spanish. Louisiana voucher program did not have ITT

estimates as it was a placement lottery. Overall effect size for programs with Bogota, Columbia removed is 0.04 (-0.00, 0.08) for three years of treatment and 0.12 (0.08, 0.16) for four or more years of treatment.

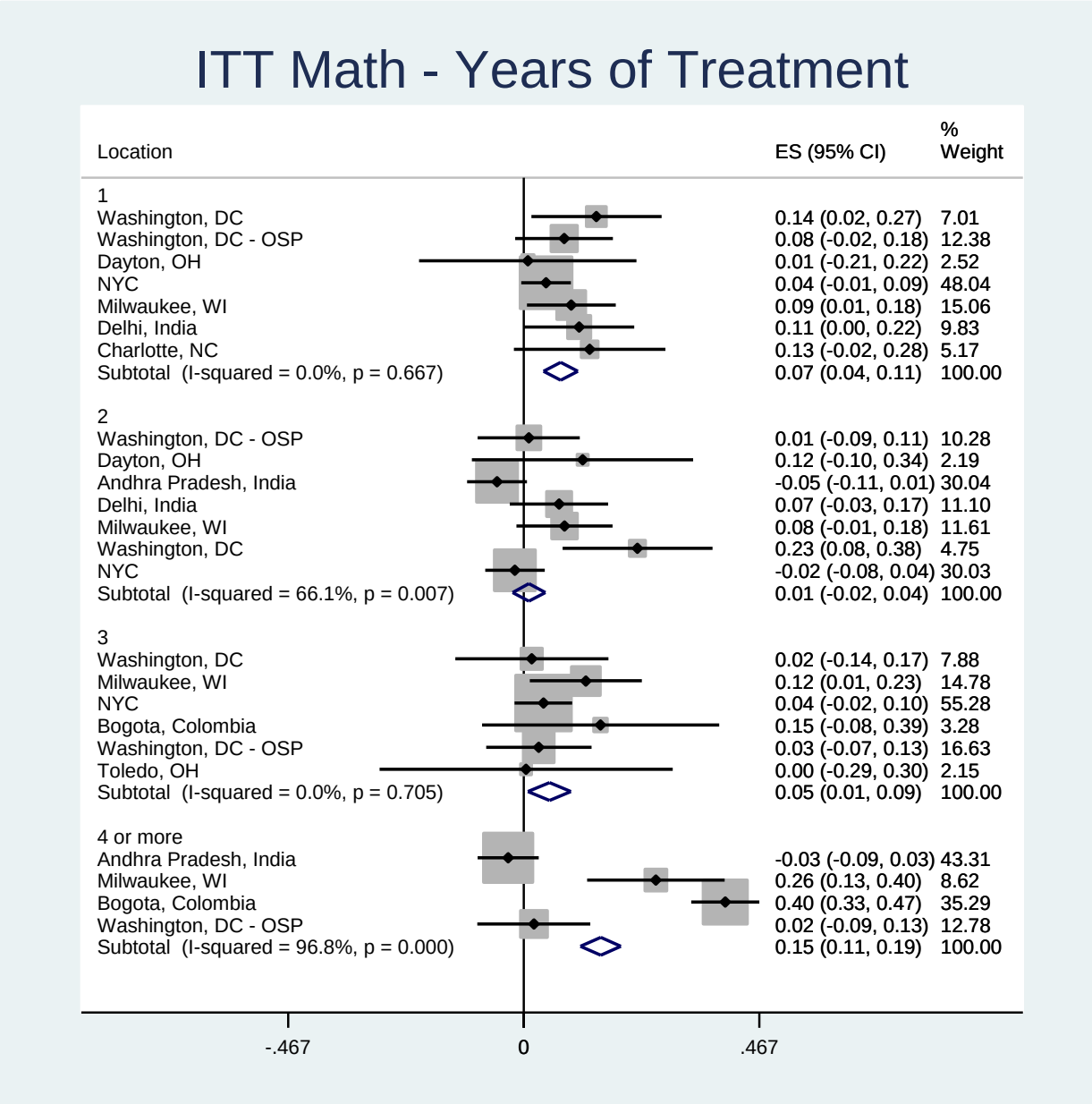
Figure 12:



Note: The Hedges' *g* estimates are based on one year effect, two year effect, three year effect and four or more year effect size calculated for each study. The boxes show overall estimates for yearly effect of programs. The grey area around each point (effect size) is the weight of each study (inverse of variance). No reading estimates are reported

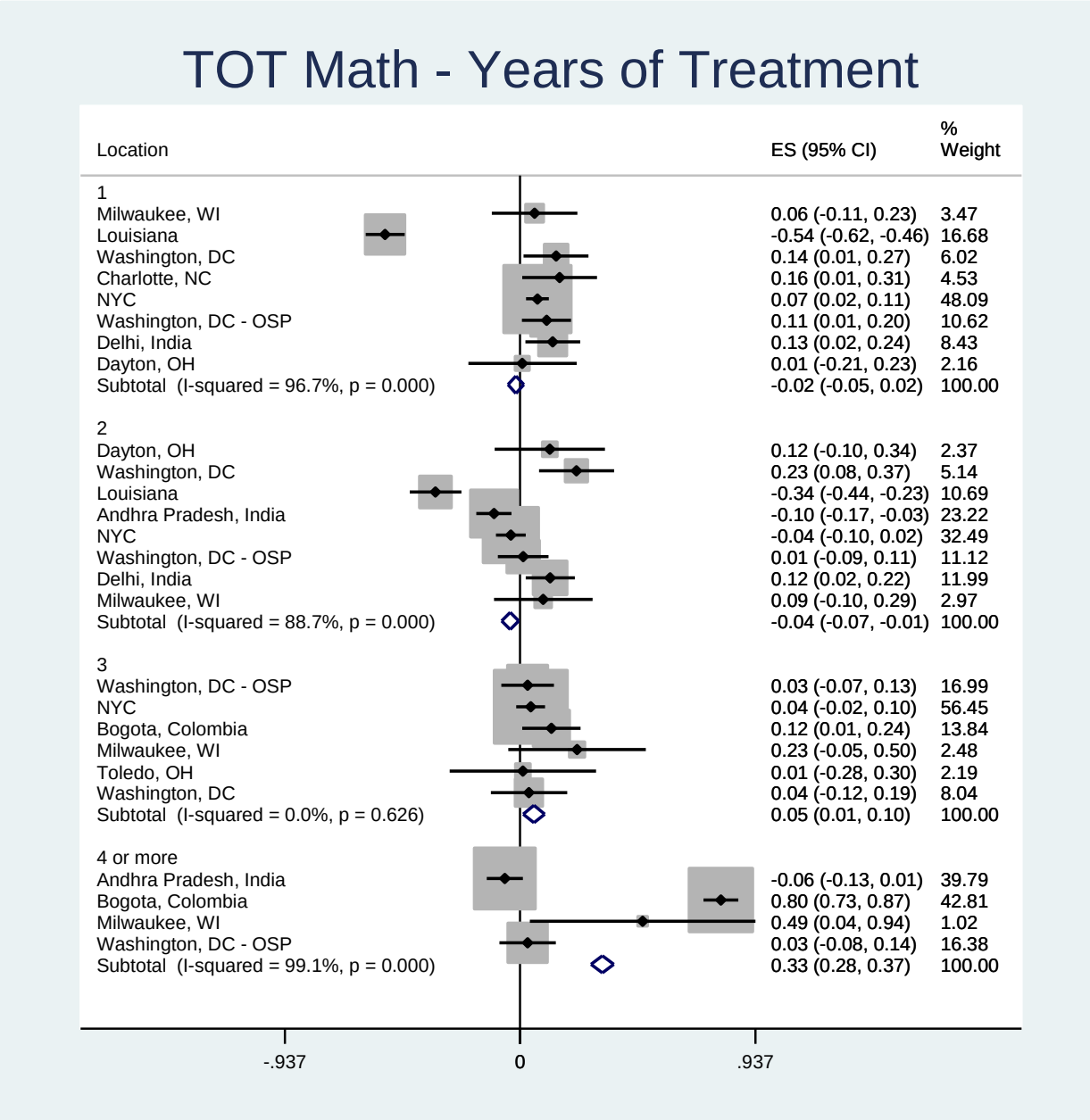
for Toledo, OH as it had only math test outcomes. Reading estimate for Delhi, India includes an overall estimate for English and Hindi. Reading estimate for Andhra Pradesh, India includes an overall estimate for English, Hindi and Telugu. Reading estimate for Bogota, Colombia is for Spanish. Overall effect size for programs with Bogota, Columbia removed is 0.03 (-0.01, 0.07) for three years of treatment and 0.25 (0.21, 0.29) for four or more years of treatment.

Figure 13:



Note: The Hedges' *g* estimates are based on one year effect, two year effect, three year effect and four or more year effect size calculated for each study. The boxes show overall estimates for yearly effect of programs. The grey area around each point (effect size) is the weight of each study (inverse of variance). Louisiana voucher program did not have ITT estimates as it was a placement lottery. Overall effect size for programs with Bogota, Columbia removed is 0.05 (0.00, 0.09) for three years of treatment and 0.02 (-0.03, 0.07) for four or more years of treatment.

Figure 14:



Note: The Hedges' *g* estimates are based on one year effect, two year effect, three year effect and four or more year effect size calculated for each study. The boxes show overall estimates for yearly effect of programs. The grey area around each point (effect size) is the weight of each study (inverse of variance). Overall effect size for programs with Bogota, Columbia removed is 0.04 (-0.00, 0.09) for three years of treatment and -0.03 (-0.09, 0.03) for four or more years of treatment.

APPENDIX A: Assumptions and Calculations, by Study

Abdulkadiroglu, Pathak, & Walters (2015): Louisiana Scholarship Program

- No ITT effects because it was a placement lottery.
- Sample attrition was 17% for lottery losers (p. 13), and Table 10 indicates the probability of observing a score is about 8 percentage points higher for lottery winners than lottery losers, so we assume 9% sample attrition rate for lottery winners. Overall sample attrition is calculated as the number of attriters divided by the assumed beginning N (1,456) where the assumed beginning N = (treatment N/(1-attrition rate of treatment group) + (control N/(1-attrition rate of control group)). Overall sample attrition, therefore, is (1,456-1,248)/1,456 = 14.3%.
- Treatment and control splits is based the following: Control group sample size is equal to the total sample size from Table 4 (1,247 in Math or 1,248 in Reading) times the loser rate from Table 10 (903/1412 or about 64%). Then the treatment group size is the Total N – Control N.

Angrist, Bettinger, Bloom, King, & Kremer (2002): Programa de Ampliacion de Cobertura de la Educacion Secundaria (PACES), Bogota, Colombia

- ITT reading effect from Table 5.
- Control group sample size from Table 2, total ITT sample size from Table 5.
- TOT sample sizes from Table 7 (Control = 562, N of "Loser Means'; Total = 1,147)

- Sample attrition (year 3) is based on 283 students who took the test (Table 2) out of the total 1,147 (Table 3).
- Program attrition: 10% from p. 1,536 and 1,547.
- TOT effects were not split into reading and math, only an overall in Table 7 (p. 1549). We calculated separate math and reading TOT estimates using the following equations:

$$TOT\ Reading = Total\ TOT * \frac{ITT\ Reading\ Impact}{ITT\ Reading\ Impact + ITTMath\ Impact}$$

$$TOT\ Math = Total\ TOT * \frac{ITT\ Math\ Impact}{ITT\ Reading\ Impact + ITTMath\ Impact}$$

Angrist, Bettinger, & Kremer (2006): Programa de Ampliacion de Cobertura de la Educacion Secundaria (PACES), Bogota, Colombia

- ITT effects for year 7 (ICFES exam scores) are the Tobit 10% results on p. 853.
- Total sample size (3,541) from footnote in Table 3. Treatment group was 58.5% of total sample size (Table 1, p. 850)
- Program attrition: 50% within three years (p. 854).
- Sample attrition Table 1 as:

$$Sample\ attrition = \frac{Full\ Sample - Obsereved\ (Valid\ ID\ and\ Age)}{Full\ Sample} = \frac{4{,}044 - 3{,}542}{4{,}044} = 12.4\%$$

- TOT effects were calculated from the ITT estimate using the following Bloom adjustment:

$$TOT\ estimate\ = \frac{ITT\ estimate}{usage\ rate}$$

where the usage rate is 1- program attrition = 1 - .5 = .5

Barnard, Frangakis, Hill, & Rubin (2003): School Choice Scholarships Foundation Program (NYC)

- Sample attrition: Utilized Table 1 for total number at randomization (676+676 = 1352), and 1,050 as the observed sample, to calculate attrition rate of 22%: (1,352-1,050)/1,352
- Program attrition: Midpoint of 20% and 27%, the percentage of children who won scholarships and did not use them (p. 301).
- ITT effects: overall estimate based on a meta-analytic average of the "Low School" and "High School" impacts presented in Table 4. "Overall" impacts (combination of different grades at application) were used.
- There was a lack of detail on sample sizes, so treatment and control group sample sizes were based on a 50/50 split of the total number of single-child families included in the analysis (p. 301).

Bettinger & Slonim (2006): Children's Scholarship Fund (Toledo, OH)

- Math effects only. ITT effect size from Table 3.
- Used some information from Bettinger & Slonim (2003) as needed.
- Sample size reported in Table 3 (N=349) was based on stacking two sets of math test scores, but this overstates the actual number of students. The footnote indicated 163

students who took both parts of the test, and 23 who took one part of the test, so we used a total sample size of 163 + 23 = 186.

- Control group is calculated as 58% of the 186 total sample, where 58% is the number of lottery losers (1,416 from p. 30), divided by the difference between the number of applicants (2,424) from p. 7 of Bettinger & Slonim (2003) and 39 "mystery winner" students who were excluded from the analysis. 58% = 1,416/(2,424-39).
- Program attrition: N/A. Table 1 on p. 30 indicates that the total number of winners was 2,385 (1,126 + 1,259). The number of losers was 1,416 (331 + 1085), but no indication of how many lottery winners actually used the vouchers.
- Sample attrition: 186 tested out of 2,385, indicates sample attrition of 92% (Table 1).
- TOT math effect was calculated from the ITT estimate using the following Bloom adjustment:

$$TOT\ estimate\ = \frac{ITT\ estimate}{usage\ rate}$$

where usage rate is 43% (p. 12).

Bitler, Domina, Penner, & Hoynes (2015): School Choice Scholarships Foundation Program (NYC):

- Sample sizes all assumed to be the same as Krueger & Zhu (2004).
- Sample attrition from Panel A of Table A1 (Bitler et al., 2015). For example, year 1 math attrition was calculated as the difference between the number of students randomized and the number of students with valid test scores (2,666 – 1,977), divided by the number of students randomized (2,666).

- Program attrition: From Panel B of Table A2 (Bitler et al., 2015). For example, in year 1, 1,022 of the 1,292 students randomized were attending a private school, indicating a first year usage rate of 79.1% and program attrition in the first year of 20.9%.
- ITT effects from Table 3, last column.
- TOT effects were calculated from the ITT estimates using the following Bloom adjustment:

$$TOT\ estimate\ = \frac{ITT\ estimate}{usage\ rate}$$

where usage rates were based on Table A2, Panel B. For example, in year 1, 1,022 of the 1,292 students randomized were attending a private school, indicating a first year usage rate of 79.1%.

Cowen (2000): Charlotte Children's Scholarship Fund

- Program attrition: 25.5% (54/212 of those offered voucher declined it), Table 1 (p. 307).
- Sample attrition: 70% based on 30% of participants with outcome testing (Table 1, p. 307).
- ITT sample sizes from Table 1.
- ITT effects from Table 2.
- TOT in this case is the Complier Average Casual Effect (CACE), the mean treatment outcome across the subpopulation of compliers.
- TOT treatment group sample size (N = 212, number of users, p. 307).

- TOT control group sample size (From Table 1: N = "Total" minus "Choice" = 347 – 158).

Greene (2000): Charlotte Children's Scholarship Fund

- Program attrition calculated as the percent of students who won but did not attend divided by the total who won (413/(413/388) = 51.6% (p. 3).
- Sample attrition: Overall sample attrition 60% (p. 3).
- TOT estimates are IV results from Table 3. T-statistic was calculated using a p-value of 0.05 and degrees of freedom of 350 (N=357 – 7 variables including constant).
- Treatment/control split was based on the ratio of Choice students to Public students in Table 2 (Choice = 145, Public is 197), applied to the total N of 357.
- ITT estimates were calculated from the TOT estimates using the following Bloom adjustment:

$$TOT\ estimate\ = \frac{ITT\ estimate}{usage\ rate}$$

Usage rate for was 48.4% (1-program attrition rate of 51.6%).

Greene, Peterson, & Du (1999): Milwaukee Parental Choice Program (MPCP)

- Sample attrition was calculated as the 1 – prob(test data available) for each group. For example, 40% of the treatment group had test data available by the third and fourth year,

so sample attrition was 60%. 48% of the control group had test data available by the third and fourth year, so sample attrition was 52%.

- Table 6 was used to calculate treatment/control splits for the ITT estimates. For example, for Reading ITT, Control N= 48/(48+63) or 43.2% of the total sample.
- TOT estimates from Table 3.
- Tables 3 and 6 was used to calculate treatment/control splits for the TOT estimates. For example, for Reading TOT, of the 758 students who had scores three of four years after application, 592 or 72% were treatment students, so the treatment N was 0.78 times the total N in table 3. For example, year three reading treatment N = 301 (from Table 3) times 0.78 (from Table 6).

Peterson, Howell, Wolf & Campbell (2003): School Choice Scholarships Foundation Program (NYC)

- ITT effects: combined African-American and Other Ethnic Group results from Table 4B.1 in Peterson, Howell, Wolf & Campbell (2003) using meta-analytic average.
- TOT effects: combined African-American and Other Ethnic Group results from Table 4.2 in Peterson, Howell, Wolf & Campbell (2003) using meta-analytic average.
- ITT treatment and Control group sample sizes in years 1 and 2 based on response rate in each year times number of vouchers offered. For example: 1st year treatment group sample size is the total number of offers times the response rate (1,300 x 82% = 1,066) from p. 195 of Howell, Wolf, Campbell, & Peterson (2002). 1st year control group sample

size is total N from Peterson, Howell, Wolf & Campbell (2003) of 1,434 minus the 1,066 treatment units.

- Response rates between treatment and control assumed to be the same according to Peterson, Howell, Wolf & Campbell (2003), p. 197, with the exception of in year 2. In year 2, the response rate was 7 percentage points higher in the treatment group than in the control group. Treatment and control split in year 2 was generated so that this differential was approximately 7 percentage points (912/1300 = 70.2% is the treatment group response rate and 284/449 = 63.3% is the control group response rate).

Peterson, Howell, Wolf & Campbell (2003): Washington Scholarship Fund (DC)

- ITT effects: combined African-American and Other Ethnic Group results from Table 4B.2 in Peterson, Howell, Wolf & Campbell (2003) using meta-analytic average.
- TOT effects: combined African-American and Other Ethnic Group results from Table 4.4 in Peterson, Howell, Wolf & Campbell (2003) using meta-analytic average.
- ITT treatment and Control group sample sizes in years 1 and 2 based on response rate in each year times number of vouchers offered. For example: 1st year treatment group sample size is the total number of offers times the response rate (809 x 63% = 510) from p. 195 of Howell, Wolf, Campbell, & Peterson (2002)
- Response rates between treatment and control assumed to be the same according to Peterson, Howell, Wolf & Campbell (2003), p. 197.
- The standard error on the year three reading impact for Other Ethnic Groups was not reported in Peterson, Howell, Wolf & Campbell (2003), but due to uniformity of standard

error patterns across years within each subject, we calculated an average. For example, the standard errors for DC reading ITT impacts for African-American students were 1.5, 1.4, and 1.5 standard deviations for years 1, 2, and 3). Reading year three ITT standard error is the average of the year one and year two standard errors (8.0 and 9.1).

Peterson, Howell, Wolf & Campbell (2003): Parents Advancing Choice in Education (Dayton, OH)

- ITT effects: combined African-American and Other Ethnic Group results from Table 4B.3 in Peterson, Howell, Wolf & Campbell (2003) using meta-analytic average.
- TOT effects: combined African-American and Other Ethnic Group results from Table 4.3 in Peterson, Howell, Wolf & Campbell (2003) using meta-analytic average.
- ITT treatment and Control group sample sizes in years 1 and 2 based on response rate in each year times number of vouchers offered. For example: 1st year treatment group sample size is the total number of offers times the response rate (515 x 56%) from p. 195 of Howell, Wolf, Campbell, & Peterson (2002).
- Response rates between treatment and control assumed to be the same according to Peterson, Howell, Wolf & Campbell (2003), p. 197.

Jin, Barnard, & Rubin (2010): School Choice Scholarships Foundation Program (NYC):

- No ITT effects, because this is just using a different TOT-methodology with the same Barnard et al. (2003) and Krueger & Zhu (2004) sample.

- TOT effects from Table 7. Same assumptions made as Barnard et al. (2003).

Krueger & Zhu (2004): School Choice Scholarships Foundation Program (NYC)

- Assumed to be same data as Bitler et al. (2015) so if statistics were not available in Krueger & Zhu (2004), we referenced Bitler et al. (2015).
- ITT treatment effects from Table 3b which uses the revised weights and without controls for baseline scores.
- For year three sample sizes, 2,770 is assumed to be the original all inclusive sample, because 1,801 was reportedly left after roughly 35% attrition. Half each of 2,770 is assumed to be treatment and control (1,385 each). Treatment and Control attrition rates (p. 638) were then used to calculate the number of treatment and control units in the analytic sample. For example 35.4% of the control group attrite, so the remaining is 895, and the remaining 906 in the total sample size are assumed to be treatment units.
- Year 1 and 2 treatment and control splits were assumed to be in the same ratio in year three.
- Sample attrition rates for each year were then calculated based on the observed sample size in a given year and the original sample size (2,770).
- Program attrition rates in each year are assumed to be the same as Bitler et al. (2015), from Table A2, Panel B.

- TOT effects from Table 6 2SLS results.
- TOT samples sizes: assumed to be the same as ITT, because not enough information.

Mills & Wolf (2016): Louisiana Scholarship Program

- No ITT effects because it was a placement lottery.
- TOT effects from IV Late estimates in fully specified model (Table 3).
- Statistics obtained directly from lead author.

Muralidharan & Sundararaman (2015): Andhra Pradesh (AP) School Choice Experiment, Andhra Pradesh, India

- ITT effects from Table VI, Panel A. Two languages impacts were meta-analyzed into one overall for year two, three impacts for year three.
- TOT effects from Table VI, Panel B. Two languages impacts were meta-analyzed into one overall for year two, three impacts for year three.
- 2 year program attrition: 39%: 39% of those offered did not use the voucher (p.10).
- 4 year program attrition: 49.2%: 39% of those offered did not use the voucher (p.10), but at the end of four years only 1,005 out of the 1,980 original treatment group were still using it. (1,980-1,005)/1980 = 49.2%

- Sample attrition rates differ by year and test but are based on Table A.2 and Table VI. For example, the year 2 English sample attrition is 14.9%: (5,316 – 4,525/5,316) where 5,316 is the sum of the 1,980 + 3,336 in Table A.2 and 4,525 is the sample size in Table VI.

Rouse (1998): Milwaukee Parental Choice Program (MPCP)

- Treatment and control group sample sizes are based on Table 1, p. 555. Assumption is that reading analytic samples are identical to math analytic samples.
- TOT effect not calculated. Overall TOT effect for Milwaukee is based only on Greene, Peterson, & Du (1999).

Wolf, Egalite & Dixon (2015): Ensure Access to Better Learning Experiences (ENABLE), Delhi, India

- Year 1 ITT: treatment-control means, difference, effect size, and p-value taken from Tables 1, 2, and 3.
- Year 2 ITT: treatment-control means, difference, effect size, and p-value taken from Table 25.2. All other statistics acquired from data output obtained directly from the authors.

- TOT effects were calculated from the ITT estimate using the following Bloom adjustment:

$$TOT\ estimate\ = \frac{ITT\ estimate}{usage\ rate}$$

where the usage rate is 0.8678.

Wolf, Kisida, Gutmann, Puma, Eissa, & Rizzo (2013): DC Opportunity Scholarship Program

- ITT reading effects from Table 3-2 and Figure 3-1.
- ITT math effects from Tables 3-2 and 4-1 and Figure 3-2.
- Program attrition: Based on p. 67-67 (year 1), p. A-34 (year 2), p. A-32 (year 3), and p. A-41 (year 4).
- TOT effects in year one and two were calculated from the ITT estimates using the following Bloom adjustment:

$$TOT\ estimate\ = \frac{ITT\ estimate}{usage\ rate}$$

where usage rates for year one and two rare based on p. 67-68.

- TOT effects in year three and four were based on percent of "never users."

Appendix B. Details on Search and Exclusion Process

	Number of Articles
Search 1 (University of Arkansas Library)	
Three library sources (EBSCO, JSTOR, ProQuest)	2,737
Duplicates Removed	-534
Unique articles (EBSCO, JSTOR, ProQuest)	2,203
Excluded Based on Title and/or Abstract	-2,075
Remaining Articles (EBSCO, JSTOR, ProQuest)	**128**
Search 2 (Google Scholar)	
Number of Google Scholar Sources Initially Found	6,706
Excluded Based on Title and Abstract	-6,549
Remaining Google Articles	157
Duplicates Removed	-9
Remaining Articles (Google Scholar)	**148**
Sum of Remaining Articles (Both Searches)	**276**
Excluded Based on Full Article	-260
Studies added through networked search	+3
Total search results (RCTs)	**19**

APPENDIX C: Formula used during meta-analysis

1. **Mean differences:** $\bar{X}_T - \bar{X}_C$
2. **SD Pooled :** $Std_{(pool)} = \sqrt{\frac{S_1^2(n_1-1)+S_2^2(n_2-1)}{n_1+n_2-2}}$
3. **Cohen's D:** $d = \frac{\bar{X}_T - \bar{X}_C}{Std_{(pool)}}$
4. **Lower bound ES (95%):** LB = $ES - SE_d * 1.96$
5. **Upper bound ES (95%)**: UB= $ES + SE_d * 1.96$
6. **Effect Size by correlation:** $ES = \frac{2r}{\sqrt{1-r^2}}$
7. **Effect Size by t ratio:** $d = t\sqrt{\frac{n_1+n_2}{n_1 n_2}}$
8. **Hedges' g (Unbiased D):** $ES(d') = \left[1 - \frac{3}{4N-9}\right] d$
9. **Standard error for effect size:** $SE_{d'} = \sqrt{\frac{n_1+n_2}{n_1 n_2} + \frac{d'^2}{2(n_1+n_2)}}$
10. **Inverse Variance (w)** $w = \frac{1}{(SE)^2}$
11. **Grand Effect size:** $\overline{ES} = \frac{\sum(w \times ES)}{\sum w}$

Where *ES* is effect size of each study, *w* is the inverse variance weight