NEPC Review: Comparing Ed Reforms: Assessing the Experimental Research on Nine K-12 Education Reforms (EdChoice, April 2020)

Reviewed by:

Bruce D. Baker
Rutgers University

May 2020

National Education Policy Center

School of Education
University of Colorado Boulder
nepc.colorado.edu

Acknowledgements

NEPC Staff

Kevin Welner
NEPC Director

William Mathis
Managing Director

Alex Molnar
Publications Director

Suggested Citation: Baker, B.D. (2020). *NEPC Review: "Comparing Ed Reforms: Assessing the Experimental Research on Nine K-12 Education Reforms."* Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/thinktank/charter-research

Funding: This review was made possible in part by funding from the Great Lakes Center for Educational Research and Practice.

NEPC
NATIONAL EDUCATION POLICY CENTER

NEPC Review: Comparing Ed Reforms: Assessing the Experimental Research on Nine K-12 Education Reforms (EdChoice, April 2020)

Reviewed by:

Bruce D. Baker
Rutgers University

May 2020

Executive Summary

A recent report from EdChoice, working with Hanover Research, identifies and reviews studies that use Randomized Control Trials (RCTs) to determine student achievement or educational attainment outcomes of nine broad "education reform" areas. The report presents counts of studies with positive, negative, and neutral findings across these areas. RCTs are presented in the report as "gold standard" studies for determining effects of specific treatments on measured outcomes. However, the report's approach has several problems and limitations. Among the RCT studies reviewed, that which is actually randomized is extremely limited and operating in a largely non-random context; the studies are not fully "randomized" or "controlled." RCT studies of charter schools tend to be limited to specific contexts, their specific models, and their particular programs and services. Private school voucher policies are similarly varied and difficult to classify as a "treatment." Moreover, as the report notes, studies categorized under the reform "open enrollment" are actually two studies involving magnet school programs. In fact, the researchers found no studies that could be squeezed into three of the reform categories. All six of the tallied reform areas—from smaller class sizes and schools to pre-k programs to school choice—score well, with positives outweighing negatives by large margins. But that truly tells us very little, and the study authors are in fact cautious in explaining their modest goals—merely tallying the quantity of research done in specific areas. The main concern with this report, therefore, is that the casual reader will take the table presenting the tallies out of context and use it to argue that charter schools and vouchers for private schools have been studied most (because they are most important) and that most of these studies find positive effects. If, however, the report is not misused in such a way, it offers a limited contribution for readers wanting to get an initial feel for the RCT research in these areas.

NEPC
NATIONAL EDUCATION POLICY CENTER

NEPC Review: Comparing Ed Reforms: Assessing the Experimental Research on Nine K-12 Education Reforms (EdChoice, April 2020)

Reviewed by:

Bruce D. Baker
Rutgers University

May 2020

I. Introduction

The policy brief *Comparing Ed Reforms: Assessing the Experimental Research on Nine K-12 Education Reforms* was prepared by Hanover Research (edited by EdChoice) and released in April of 2020.[1] EdChoice.org describes itself as "a nonprofit, nonpartisan organization dedicated to advancing full and unencumbered educational choice as the best pathway to successful lives and a stronger society. EdChoice believes that families, not bureaucrats, are ed best equipped to make K–12 schooling decisions for their children." Hanover Research describes itself as "a brain trust designed to level the information playing field." The brief uniquely states in its header that the reviews of studies were conducted by Hanover Research and the coding of studies and editing of the report by EdChoice.

The goal of the policy brief is to summarize the body of "experimental" research on nine identified "education reform" areas, where "experimental" specifically meant studies that involved some degree of random assignment of students to treatment and control groups (Randomized Control Trials, or RCTs). Studies were not necessarily restricted to studies that had been subjected to peer review. Studies were restricted to studies that measured outcomes in terms of academic achievement or educational attainment. The authors explain their goals as follows:

> We reviewed experimental research on these nine education reform areas not to say one reform is "better than another," but to report that we know more or less about certain reforms' effects compared to others based on the volume of existing and reviewed experiments. Our goal in presenting this research is not to compare these reforms or to promote one improvement approach over another. We wanted to find out what has been rigorously studied and where there are needs and opportunities for high-quality empirical research (p. 1).

The authors provide brief summaries of the collective bodies of studies in each of the following nine reform strategies: a) class size reduction, b) common enrollment applications, c) open enrollment choice programs, c) portfolio management models, d) pre-kindergarten programs, e) private school vouchers, f) charter schools, g) school size (small schools) and h) school takeovers. The authors identify three of these areas—common enrollment applications, portfolio management, and school takeovers—as having no identifiable experimental studies.

After tallying the available studies across the identified categories, the authors conclude only that:

> Our overall punchline is not new, but we believe it is still very important: We need more high-quality research in all of these reform areas. As policymakers and advocates continue to innovate and implement new programs aimed at fostering K–12 student success, we must continue to set goals and study the outcomes so that we can determine whether we are, in fact, succeeding (p. 2).

II. Findings and Conclusions of the Report

Below is a copy of the summary table, presumably tallied by staff from EdChoice based on the study reviews provided by Hanover Research. This summary table follows a relatively common format for tallying research either on a specific topic or, in this case, a set of related topics. Some authors choose to tally counts of specific "estimates" of effects (where some studies include more than one test of a difference between treatment and control groups), while others, like this, count the number of studies. In typical style, the table then summarizes the number of studies that found some positive effect, some negative effect, or no visible effect. Neither the table nor the narrative summaries of the studies focus extensively on the size of the effects identified. That is, they set aside the question of whether the effects are large enough to be of importance. The narrative summaries nonetheless provide useful context and descriptions of the studies tallied.

Table 1. Replication of Summary Table from EdChoice Report (p. 4)

Reform Type	Total Count of Studies	Any Positive Effect	No Visible Effect	Any Negative Effect
Public Charter Schools	22	21	1	0
Private School Choice (Vouchers)	21	13	6	3
Pre-Kindergarten	8	6	1	2
Class Size (Small Classes)	4	4	0	0
School Size (Small Schools)	3	3	0	0
Open Enrollment	2	2	0	0

Common Application/Unified Enrollment	0	0	0	0
Portfolio Management	0	0	0	0
School Takeover	0	0	0	0

Notes: One charter school Randomized Control Trial (RCT) focused on voting as the examined outcome. That study is not included in the displayed count.

A number of RCTs have assessed private school choice program effects on non-academic outcomes: civic values and social behaviors (7 studies); parent satisfaction (5 studies); racial/ethnic integration (2 studies)

One class size RCT study has led to many follow-up RCT studies. Those studies considering outcomes other than student achievement or educational attainment are not reflected in this table. Because Nye et al (2001) and Pate-Pain et al (1997) analyze the same program and overlapping observed outcome, we consolidated those two RCTs into a study count of one.

Two open enrollment RCTs could also be categorized as a [sic] public magnet school studies.

Two RCT studies analyzing common/unified enrollments systems did not examine student achievement or attainment.

The modest goal of their research and resulting brief was to identify the quantity of experimental research (RCTs, or presumably "gold standard") that had been conducted on each of these reforms. How the reforms were selected is not stated. To that end, they conclude: "We need more high-quality research in all of these reform areas" (p. 2).

III. The Report's Rationale for Findings and Conclusions

The purported goal of the brief is to address how much researchers have studied certain reforms using RCTs. The authors do not claim or attempt to draw any significant conclusions regarding study findings (i.e., what actually works better), and they explicitly state as much.

There are few policy questions that can be sufficiently answered by simply tallying counts of studies and positive and negative results of studies. What gets studied is not random. Where it gets studied is not random, and findings may vary from one location to another. Merely categorizing a study as having "positive" or "negative" findings omits valuable information about effect sizes. And a given intervention may yield multiple published studies, adding to the tally for that policy approach but telling us nothing about other instances or locations where that policy approach has been used.

The one question that can be answered by such a tally method is the primary one the authors set out to answer—which is, how many randomized studies had been published on each of these topics. That said, it requires significant confidence to infer that you've actually captured the universe of studies that might relate to any of these broad topics.

The authors identified and reviewed many of the best-known studies of the topics assigned, from studies of KIPP's Lynn, Massachusetts charter school, to studies of Harlem Children's Zone (HCZ) charter schools in New York, to studies of the Milwaukee and Washington, DC voucher programs, to the Tennessee class size and New York City small schools studies.

The question posed by the authors was simple, and the method used was relevant for an-

swering that question. The bigger issue is how useful that question is either from a policy perspective, or for guiding future research intended to inform policy.

IV. The Report's Use of Research Literature

The authors selected studies based on the assigned reform strategies and limited to specific methodologies. They focused their review on studies they characterized as "experimental" or Randomized Control Trials, rationalized below:

> The best methodology available to researchers for generating "apples-to-apples" comparisons is a randomized control trial (RCT), which researchers also refer to as random assignment studies or experimental studies. Essentially, these studies conduct experiments—with treatment and control comparisons—and are widely considered to be the "gold standard" of research methods. We prefer evaluating school choice programs and other education reforms based on experiments and limit the scope of this review to assessing only RCTs (p. 1).

The authors do not address the many limitations of RCTs either generally[2] or for guiding policy and practice in the social sciences.[3]

The authors further explain that the studies identified fall into two categories regarding how they measure "treatment" effects on student outcomes (p. 1):

> Intent-to-treat (ITT) effects, which compares outcomes between students who won the lottery and students who did not win the lottery. ITT is the estimated effect of being chosen for treatment via randomization.
>
> Treatment-on-the-treated (TOT) effects, which compares differences in outcomes between students who attended a private school and students who did not attend private school, regardless of their lottery outcome. TOT is the estimated effect of enrolling or participating in a given reform/policy/program, hence receiving the treatment.

Nine education reform areas were selected for the brief. As described in the brief (p. 1):

> EdChoice partnered with Hanover Research to find out what research has been conducted in nine major education reform areas focusing on outcomes related to student achievement or education attainment:
>
> - class size (small classes)
> - common enrollment applications/unified enrollment systems
> - open enrollment (inter-/intra-district)
> - portfolio management
> - pre-kindergarten
> - private school choice

- public charter schools
- school size (small schools)
- school takeover

No rationale is provided for the selection of these "reform" areas.

The brief provides a summary table with notes (Table 2, on p. 6 of the brief) that explains that private school voucher studies and charter school studies are primarily based on lottery selection mechanisms. Further, that studies of "open enrollment" programs, broadly classified, were more specifically studies of magnet schools in Connecticut and North Carolina. The class size studies reviewed all originate from the Tennessee STAR studies.

The authors provide reasonable descriptions of the studies reviewed, their methods, measures and context even in the brief summaries provided in the policy brief, which notes that more thorough summaries are to be available online (link not yet included in brief).[4] The brief provides a map to identify the contexts of the studies included, and it clarifies along the way, for each reform category, the relevant contexts. The brief provides detailed notes (with tables, such that the tables can stand alone) regarding studies that were excluded and why.

The methods and language used in the report also seem sufficiently transparent about the roles of EdChoice and Hanover staff in preparing the document. For example, the section on charter school studies starts, "Hanover Research identified 22 RCT studies of public charter schools that report effects on students' achievement or educational attainment," and the section on studies of private school vouchers starts, "EdChoice has identified 21 experimental studies reporting the effects of private school choice (voucher) programs on participating students' test scores and educational attainment." EdChoice publishes a separate report that includes a presentation of these studies.[5]

V. Review of the Report's Methods

As noted above, the authors select what they refer to as "experimental" design, or RCT studies of nine reforms, and then apply what I often refer to (with no offense to the authors) as a "bean count" method to determine how many of these "gold standard" studies have been done on each of the nine reform areas. The authors find that the largest number of studies have been conducted on charter schools and private school voucher programs and that no such studies have been conducted on "portfolio management," "common application/unified enrollment" systems, or "school takeovers." The authors conclude that more research is needed across these reform areas. The method is suitable to the findings. But the findings are limited in their value for guiding policy or policy researchers.

On "Treatments" and "Randomization"

The authors succinctly and accurately describe *Intent to Treat* and *Treatment on the Treated* effects, as they relate to the various studies they reviewed. The type of treatment effect

being evaluated is a second-level issue, however, that presumes a sound foundation. Understanding the concept of an RCT, or scientific experimentation and its application to social science contexts, requires a step back. The goal of an "experimental" design is to be able to isolate the effect of a specific "treatment" on a specific "outcome" or set of outcomes. That is, to determine that the treatment is what affected the outcomes, holding all other conditions constant.

Ideally, a finding of a positive treatment effect in one context, or at small scale, might be used as a basis for testing the treatment in other contexts or scaling up that treatment, assuming two things. First, that we know what the treatment is—what are the elements of that treatment—and second, that the outcomes affected by the treatment are outcomes we really want to affect elsewhere or more broadly. These descriptions may seem rather basic and unnecessary, but they are important when considering the usefulness of an otherwise innocuous classification, summary and tally of research findings on education reforms.

Most of the reforms, or reform categories identified for review in this brief, are not "treatments," per se. Because the specifics of the various reform initiatives are not defined, they are not any specific thing, making for a particularly awkward research task. This undoubtedly helps explain why the researchers didn't find experimental research on the three assigned categories that are especially difficult to define as "treatments" for experimental design purposes.

Among the nine reforms, assigning students to smaller vs. larger class size is perhaps the clearest example of an identifiable "treatment," where class size ranges were specified for both treatment and control groups. That is, one factor, class size, is varied while others are held constant for both control and experimental groups. Perhaps next in line are studies of pre-kindergarten programs where some students are provided an intervention that others are not. Yet there is more variation and ambiguity among the control group in studies of this type. The treatment evaluated should be sufficiently precise for purposes of replication and/or scaling up. That is, it's important to know what the treatment really is.

Most of the studies reviewed were of *charter schooling* or *vouchers for private schooling*, neither of which are "treatments."[6] These are governance structures that may or may not relate to the provision of specific identifiable, replicable or scalable treatments. Drawing relevant policy conclusions requires far more detailed information about what was included as the treatment. What programs, strategies and resources were provided by the charter school or private school in comparison to the control group? If the study is of a specific charter school or schools under a specific charter operator with a specific model, the treatment is that specific model, not "charter schooling" per se, and the "control" group is not some generalizable "district schooling" model but the array of specific schools into which lotteried-out students were placed.

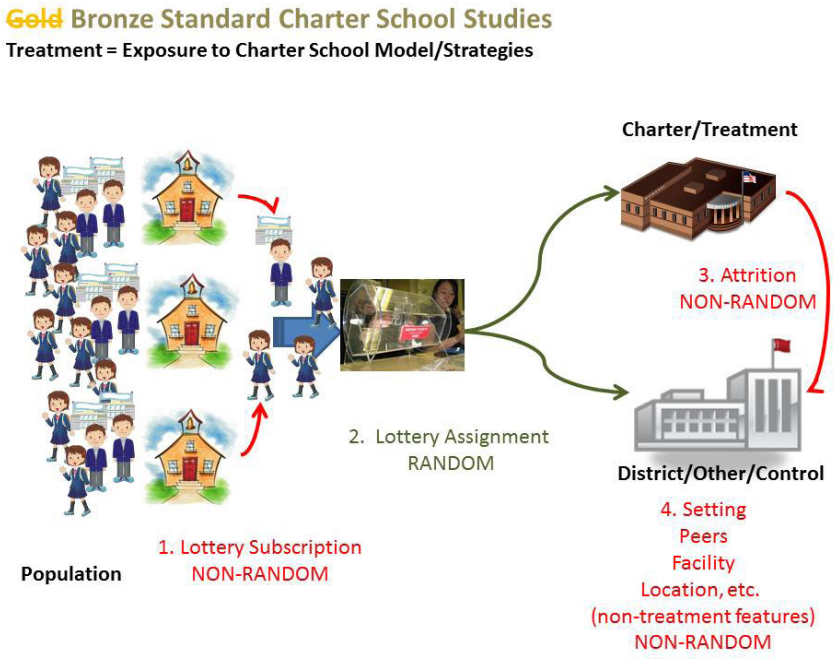
The authors of the brief are cognizant of these ambiguities and point out specific studies that involve "no excuses" charter schools, and studies attempting to parse in greater detail the underlying strategies used by those schools.[7] But labeling a charter school as "no excuses" to characterize a model adopted by that school is only marginally less ambiguous—unless the goal is to discern whether merely declaring a school as "no excuses" results in different outcomes (though this would actually require randomly assigning that label to schools that

do, and do not subscribe to supposed "no excuses" approaches). The counterfactuals in these studies are more ambiguous and varied than the treatment.

It is also important to understand that "lottery-based" studies achieve only limited randomization, and should perhaps be called LRSs—Limited Randomization Studies (dropping the "C" as well because they do not involve true control groups). Certainly, even limited randomization has benefits over none. But lottery-based studies fall well short of actual randomized experiments. The vast majority of studies included in the brief are "lottery-based" studies of participants in private school voucher programs or in charter schools. The studies that were classified under "open enrollment" programs are similar studies of children attending magnet schools. Even the New York City small-schools studies were "lottery-based" studies of schools of choice.

Figure 1 (drawn from a 2012 blog post[8]) illustrates that lottery-based assignment studies for charter schools only randomize one step in the process. Those who enter the lottery for specific schools are a non-random subset of families. They are presumably motivated to attend the charter school under investigation. A subset is randomly selected for the charter school as an entering cohort. That's the only randomized step in the process. Presumably, those who entered but lost in the lottery were similarly motivated, thus more comparable than those who didn't enter the lottery at all. Some, among those lotteried-in to charters, choose to leave. That part is non-random, as is where they end up. Some who are lotteried out may leave the district, attend private schools or other charters, while others will attend district schools (regular or magnet) where they may be tracked as part of the "control" group, along with other students (peers) who did not enter the lottery at all. But all of this sorting is non-random. And the comparison group is hardly monolithic, within any study or across contexts and studies. Perhaps most importantly, "charter schooling" simply isn't a specific "treatment."

Figure 1



To summarize, most of the studies reviewed were lottery-based studies of private school choice or charter schools. Importantly, "lottery-based" studies aren't really randomized, and "charter schooling," "private schooling," or "magnet schooling" aren't really "treatments." The various counterfactuals in these studies aren't a monolithic (or even completely random) "control" group. This is not to say these studies can't provide useful insights and, in fact, different insights from studies that have no randomization. Researchers engage in such studies because in most cases, we simply cannot (nor should we) experiment (in a more precise and accurate sense) with the lives of actual children. For purposes of this review, these concerns are relevant to the above-quoted "The best methodology available…" passage from the publication. It would have been helpful for readers if the brief included a discussion of some or all of these limitations of the RCTs (LRSs) that it relies on.

Do the Studies Reflect the Broader Context and Reforms?

The distribution of academic research in education rarely accurately or thoroughly reflects the distribution of what's actually happening (or should be) in our education systems, across contexts and children. Rather, it more often reflects the whims and preferences of those who fund research on education reform, and is often framed and conducted according to the data and measures that are most readily available. The EdChoice policy brief reinforces this fact. During a period of great policy interest in charter schooling, and increased outcome measures available for children lotteried into and out of oversubscribed charter schools in certain settings, researchers produced a plethora of studies comparing the outcomes of the lotteried-in and lotteried-out children, largely accepting "chartering" in and of itself as a form of treatment. As time went on, few scratched beneath that surface.

The authors do well at providing information on the context of the various studies reviewed. What is revealed through these descriptions is that certain types of programs and services have only been studied in certain contexts. And in some cases, when a reform or program is studied across different contexts we get different findings, as with voucher programs which have revealed positive effects in some contexts, and negative effects elsewhere. This may either be because the treatment group actually did better, or because they were compared against a stronger or weaker control group. It's easier for charter schools or private schools to compare favorably against "district" schools where district schools are weaker, and vice versa. It's all relative. It's all context-sensitive. Which complicates application to other, different contexts.

It is questionable to project findings from studies of specific charter school operators in Massachusetts or New York onto the charter sector as a whole, as "charter schooling" per se, varies widely and is not a clearly identifiable "treatment." The authors do not make such projections, and describe the studies and their contexts accurately. In fact, the brief appropriately describes the specifics of the types of schools studied under the broader assigned reform categories, and points out that the studies included under the category "open enrollment" are studies of random assignment to district operated magnet schools.

But for the reader who looks only at Table 1, the summary suggests that there are lots of rigorous studies on "charter schools" as if it's a treatment, and that "charter schools" yield

positive outcomes. The logical conclusion? We must have more charter schools, or at least more kids in them! But the listed studies are studies of specific charter schools, providing specific types of programs and services with specific resources in specific contexts. From a policy standpoint it would be more valuable to figure out how to replicate the programs, services and resources of the schools found to be effective than to simply increase the number of charter schools. From a research standpoint, it may be more useful to figure out how to isolate the programs and services provided by these schools, and replicate those across other schools (district, charter, private) using more thorough RCT designs to test their efficacy.

Table 1 also suggests that the two studies of "open enrollment" as a "reform" show positive effects, and none show negative effects. The authors rightly point out that the studies placed in this box are actually studies of interdistrict magnet school programs. The magnet schools (their programs, services, etc.) are the treatment, not "open enrollment." Adopting an interdistrict open enrollment policy in the absence of well-funded, high-quality magnet schools is unlikely to yield similar effects.

VI. Review of the Validity of the Findings and Conclusions

The authors reasonably conclude:

> Our overall punchline is not new, but we believe it is still very important: We need more high-quality research in all of these reform areas. As policymakers and advocates continue to innovate and implement new programs aimed at fostering K–12 student success, we must continue to set goals and study the outcomes so that we can determine whether we are, in fact, succeeding (p. 2).

Again, the stated goal of the brief is to address how much we have studied certain reforms. It does not purport to draw any significant conclusions regarding study findings—e.g. what actually works better.

To that end, this is the question that can be answered with a tally method.

VII. Usefulness of the Report for Guidance of Policy and Practice

This report, while not taking any problematic leaps or making hugely problematic assumptions, falls flat in terms of being useful for guiding policy, practice, or research. The report does provide a concise explanation of studies of two types of "treatment effects" commonly evaluated in quantitative social science research on schools and school systems, along with reasonable descriptions of many of those studies, citing more thorough descriptions to be posted online to accompany the brief. To that end, the brief does what it set out to do.

But, the study concludes from its tally that more research—specifically RCTs—need to be

done on many areas of education reform, setting aside the distinct possibility that the reason there are no RCTs of "reforms" like unified enrollment systems, school takeovers, or portfolio management is that these are not specific treatments that even could be plausibly studied by RCT. Note that nowhere in the brief is a description of exactly what is meant by any of these nine reform classifications. What are the supposed "treatments" to which children would be randomly assigned (one being the assignment system itself)?

My greatest concern is that the casual reader and user of such research will use Table 1 out of context to argue that charter schools and vouchers for private schools have been studied most (because they are most important) and that most of these studies find positive effects. Therefore, we should promote spending more public dollars to send more kids to charter schools and voucher schools—schools that in other contexts may look nothing like the specific schools studied and found effective, by the narrow available measures and under the unique conditions presented in these studies.

Finally, the brief operates under the narrow assumption that RCTs (of which most of the studies tallied really are not) are of the greatest value in guiding policy for the improvement of schools and school systems. While RCTs offer precision in sorting out treatment effects (assuming "treatment" is precisely identified), practical and ethical concerns limit researchers' ability to conduct true RCTs on education reforms. Further, many complex, holistic reform strategies can only be fully understood when adopted across many, varied contexts, and evaluated thoroughly through a mix of quantitative and qualitative strategies.

Notes and References

1 EdChoice (2020), *Comparing ed reforms: Assessing the experimental research on nine K–12 education reforms*, retrieved April 27, 2020 from https://www.edchoice.org/wp-content/uploads/2020/04/comparing-ed-reforms.pdf

2 Grossman, J., & Mackenzie, F.J. (2005). The randomized controlled trial: gold standard, or merely standard?. *Perspectives in biology and medicine*, *48*(4), 516-534.

3 Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2-21.

Thomas, G. (2016). After the gold rush: Questioning the "gold standard" and reappraising the status of experiment and randomized controlled trials in education. *Harvard Educational Review*, *86*(3), 390-411.

4 On page 5, the authors explain: "More in-depth summaries of the specific studies' findings can be accessed in a supplement to this brief posted online at [URL HERE]. In that supplemental and technical report, when applicable, Hanover includes findings from the U.S. Department of Education's What Works Clearinghouse (WWC), which reviews individual education research studies for quality, impact, and implications."

5 EdChoice (2020), *The 123s of school choice: What the research says about private school choice programs in America*, 2020 edition, retrieved April 27, 2020 from https://www.edchoice.org/wp-content/uploads/2020/04/123s-of-School-Choice-2020.pdf

6 I provide a discussion of "vouchers" as a "treatment" in this blog post: https://schoolfinance101.wordpress.com/2012/08/24/helicopters-can-improve-minority-college-attendance-other-misguided-policy-implications-comments-on-the-brookings-voucher-study/

7 I provide a critique of those studies here: https://schoolfinance101.wordpress.com/2012/01/26/beneath-the-veil-of-inadequate-cost-analyses-what-do-roland-fryers-school-reform-studies-really-tell-us-if-anything/

8 https://schoolfinance101.wordpress.com/2012/12/20/thoughts-on-randomized-vs-randomized-charter-school-studies/