A Serious Statistical Mistake in the CREDO Study of Charter schools

Caroline M. Hoxby
Stanford University and NBER
August 2009

Abstract

A recent study of charter schools' effect on student achievement has been published by CREDO (2009). It contains a serious statistical mistake that causes a negative bias in its estimate of how charter schools affect achievement. This paper explains that mistake. Essentially, the achievement of charter school students is measured with much more error than the achievement of the controls, which are not individual students but are group averages of students in the traditional public schools. By using the achievement data as both the dependent variable and (lagged) an independent variable, the CREDO study forces the estimated effect of charter schools to be more negative than it actually is. This paper notes that the CREDO study violates four rules for the empirically sound use of matching methods to evaluate charter schools' effects. The main conclusion is that the CREDO study is not reliable, most obviously because the statistical mistake means that its estimates of the charter school effect are substantially biased downwards from the truth.

I. Introduction

A recent study by CREDO (2009) attempts to estimate the effect that charter schools have on their students' achievement. The study contains a serious statistical mistake that causes a negative bias in its estimate of how charter schools affect achievement. This brief paper explains that mistake and also describes how the CREDO study violates several rules for the empirical sound use of matching methods to evaluate charter schools' effects.

The statistical mistake in the CREDO study arises because charter school students' achievement is systematically measured with much more measurement error than control "students" achievement. The word "students" is in quotation marks in the previous sentence because each charter school student is matched to a *group* of traditional public school students whose *average* achievement is used. That is, the group average is treated as though it were an individual student. Of course, the averaging makes measurement error much smaller for the controls than for the charter school students. A key right-hand-side covariate in the regressions is the "student's" prior test score, which should have a different coefficient for charter school students than traditional public school students owing to the systematic difference in measurement error. Because the regression restricts the estimated coefficient to be the same for both types of "student," the charter school effect is biased downwards. The downward bias in the estimated charter school effect is shown in equations below. It is of substantial magnitude, so the mistake is a serious one.

II. The basic empirical strategy of the CREDO study

The CREDO study does not have data on charter schools' admissions lotteries, so it does not use a randomization-based method of evaluation. Randomization is the "gold standard" method of evaluating charter schools' effects on students' achievement because it effectively eliminates all forms of selection bias so long as (i) randomized admissions lotteries were used and (ii) a sufficient number of students participated in them.

Lacking lottery data, the CREDO study depends on a matching method based on charter school students' prior histories in the traditional public schools (hereafter, "TPS-history matching"). The CREDO matching has a number of problems that are described in Section IV below. However, it is not matching *per se* that causes the serious statistical problem in the paper.

The CREDO study does not match individual charter school students to individual traditional public school (TPS) students with similar demographic characteristics. Instead, it matches each charter school student to a *group* of students in traditional public schools. A charter school student can potentially be matched to a group that contains many students. The study then computes average achievement and other average characteristics of each *group*. Thereafter, the study treats these group averages as though they were students.[1]

The group for each charter school student is selected according to the following procedure. Each charter school is associated with a set of traditional public schools based on which schools their students attended *before* they applied to the charter school. Naturally, this information is not available for many charter school students because they applied as kindergarteners, previously attended a school outside the local area, previously attended a private school, or simply do not have this information recorded. Nevertheless, a set of traditional public schools is picked for each charter school.

Then, each charter school student is matched to all students from his charter school's set of

[1] The CREDO study does compute standard errors that are robust, but it apparently does not cluster by student or otherwise allow for the fact that many of its observations are not independent. If it were to compute appropriate standard errors, they would be larger.

traditional public schools who:
(i) are of the same gender;
(ii) are of the race and ethnicity;
(iii) are in the same grade;
(iv) participate in special education or not (a match on the particular special education code is not required);
(v) similarly participate or fail to participate in the national school lunch program;
(vi) are categorized similarly as an English Learner or not (a match on language is not required);
(vii) have a test score from the previous year that is within 0.1 standard deviations of the charter school student's test score.

It should be noted that the test score range for the matches is quite wide. Since students in a school are usually clustered within about 0.4 of a standard deviation in test scores, a charter school student is potentially being matched *on scores* to one half of students in the set of traditional public schools. Of course, he may be matched to fewer students owing to his having an unusual race for the area, an unusual ethnicity, and so on. Unfortunately, the CREDO technical report does not contain any information on the number of students in the matched groups. (It should.) However, since each charter school is matched to multiple traditional public schools and since these traditional public schools are likely to have students who are fairly similar socio-demographically to the charter school, it is likely that most charter school students are matched to a group that contains numerous students.

III. The statistical mistake

It is well known that tests measure a student's achievement with error. Suppose that each student's true achievement is A_{ijt} where A is for achievement, i indexes the individual student, j indexes his school, and t indexes the school year. Suppose that ω_{ijt} is the measurement error. Then, we do not actually observe a student's *true* achievement, we observe:

$$\tilde{A}_{ijt} = A_{ijt} + \omega_{ijt} \quad \textbf{(1)}$$

which is an erroneous measure. For the sake of argument, let us assume that the measurement error is as benign as measurement error can be. That is, it is classical measurement error such that:

$$mean(\omega_{ijt})=0 \ , \ cov(\omega_{ijt},A_{ijt})=0 \ , \ cov(\omega_{ijt},\omega_{ij,t-1})=0 \ , \ cov(\omega_{ijt},A_{ij,t-1})=0. \quad \textbf{(2)}$$

Elementary statistics (the Law of Large Numbers) tells us that the measurement error will be much smaller for the group average observations than for the observations of individual charter school students. The CREDO study ignores this fact and treats the group average observations as though they were individual students. This generates a very substantial negative bias in the CREDO study's estimate of the effect of charter schools. To see this, some straightforward statistical math is required. In the equations below, assume that, without loss of generality, the measurement error is zero for the control group average "students" and non-zero for the charter school students. (In fact, as we shall see, all that is required is that there is *less* measurement error in achievement for the control "students.")

The CREDO study asserts that the model of achievement is as follows:

$$A_{ijt} - A_{ij,t-1} = \mu_0^{TPS} + \mu_1 A_{ij,t-1} + \varepsilon_{ijt} \quad \textbf{(3)}$$

for students in the traditional public schools (TPS) and

$$A_{ijt} - A_{ij,t-1} = \mu_0^{charter} + \mu_1 A_{ij,t-1} + \varepsilon_{ijt} \quad \textbf{(4)}$$

for students in charter schools. (Without loss of generality, other covariates on the right-hand side of the equations have been omitted for simplicity.)

The specification given by equations (3) and (4) is not a completely standard specification for studies of charter schools, but it is quite plausible. It asserts that charter schools affect a student's achievement gain over his previous year's test score. That is, the CREDO study's goal is to estimate $\mu_0^{charter}-\mu_0^{TPS}$.

Suppose that there is no measurement error in achievement for the control (TPS) "students", who are actually group averages. Thus, if we were estimate equation (3) solely for control "students," we would obtain a consistent estimate of μ_1:

$$\hat{\mu}_1 = \frac{cov(A_{ijt}-A_{ij,t-1},A_{ij,t-1})}{var(A_{ij,t-1})} \Rightarrow \mu_1 \qquad \textbf{(5)}$$

where $\Rightarrow$ stands for "asymptotically convergences to."

Charter school students' achievement is measured with classical measurement error as defined above:

$$\tilde{A}_{ijt} = A_{ijt}+\omega_{ijt} \quad , \quad \tilde{A}_{ij,t-1} = A_{ij,t-1}+\omega_{ij,t-1} \ . \qquad \textbf{(6)}$$

Thus, if we were estimate equation (4) solely for charter school students, we would obtain:

$$\breve{\mu}_1 = \frac{cov(A_{ijt}-A_{ij,t-1},A_{ij,t-1})}{var(A_{ij,t-1}) + var(\omega_{ij,t-1})} - \frac{var(\omega_{ijt})}{var(A_{ij,t-1}) + var(\omega_{ij,t-1})} \qquad \textbf{(7)}$$

which is not consistent for μ_1.

For convenience, call the bias factor ζ:

$$\zeta=\breve{\mu}_1-\mu_1 \ . \qquad \textbf{(8)}$$

The CREDO study combines data on control "students" and charter school students and thereby estimates:

$$\begin{aligned}&(A_{ijt}-A_{ij,t-1}) + (\omega_{ijt}-\omega_{ij,t-1})\cdot I_{ijt}^{charter} = \\ &\mu_0^{TPS} + (\mu_0^{charter}-\mu_0^{TPS})\cdot I_{ijt}^{charter} + \mu_1 A_{ij,t-1} - \zeta A_{ij,t-1}\cdot I_{ijt}^{charter} + \varepsilon_{ijt}\end{aligned} \qquad \textbf{(9)}$$

or (rewriting);

$$\begin{aligned}&(A_{ijt}-A_{ij,t-1}) + (\omega_{ijt}-\omega_{ij,t-1})\cdot I_{ijt}^{charter} = \\ &\mu_0^{TPS} + [\mu_0^{charter}-\mu_0^{TPS} - \zeta A_{ij,t-1}]\cdot I_{ijt}^{charter} + \mu_1 A_{ij,t-1} + \varepsilon_{ijt} \ .\end{aligned} \qquad \textbf{(10)}$$

By estimating the term in square brackets instead of the charter school effect it intended to estimate ($\mu_0^{charter}-\mu_0^{TPS}$), the CREDO study estimates a negatively biased effect. The bias is given by:

$$-\zeta A_{ij,t-1} \qquad \textbf{(11)}$$

There are two separate reasons for the negative bias. First, the estimated charter school effect is biased toward zero because the denominator of the first term on the right-hand side of equation (7) is larger than the denominator of the right-hand side of equation (5). The denominator is larger because $var(\omega_{ij,t-1})$ is added. Second, the second term on the right-hand side of equation (7) is obviously negative.

It will be seen that the CREDO study produces a negatively biased estimate of the charter school effect so long as the measurement error is larger for the charter school students than for the control group average "students." This is because all that is required is that the estimate shown in equation (7) is more negatively biased for charter school students than for group average "students." This will be true so long as the variance of measurement error ($var(\omega_{ij,t-1})$) is larger for charter school students than for control "students." This is of course true since the group averaging reduces the variance of the measurement error for the control "students".

The bias is serious because we know that measurement error or some similar phenomenon causes the coefficient on $A_{ij,t-1}$ to be considerably smaller than 1 in a typical estimation of an equation such as (4) that is based on individual student data. An estimate of 0.7 to 0.8 is common. This suggests that the bias factor, ζ, must be at least 0.2 just from the *first* source of bias. The second source of bias will add another 0.2 (based on backing out the $var(\omega_{ij,t-1})$ that would produce estimates of 0.7 to 0.8, as mentioned). This means that the bias is at least 0.4 times achievement lagged once ($A_{ij,t-1}$). **This is bias of very substantial magnitude.**

IV. Other Problems with CREDO's Matching-Based Method

If we are to understand the frailties of matching-based methods for estimating charter school effects, it helps to back up and think briefly about the "experiment" that we would like to discern in non-experimental data. We would like to find circumstances in which one in each pair of identical students is randomly assigned to attend a charter school and the other is randomly assigned to attend traditional public school. People often read the previous sentence and assume that the key word is "identical," whereas the key word is "random." Small departures from students' being identical will likely produce small differences in achievement because achievement is a fairly continuous function of most continuous student characteristics. Thus, whether students are identical or merely very similar may not matter much. On the other hand, once assignment is based on self-selection instead of a random variable, we have no way to determine how biased the resulting estimate is. Since the bias is based on variables we do not observe, what may *seem* like a small departure from randomness can generate bias of considerable size.

The fundamental problem with using matching-based methods is that, in the absence of a lottery that forces similar students to attend different schools, we know that the "nearly identical" students are not in fact nearly identical *since they decided to make different choices*. They may be nearly identical on a few observable variables, but they are evidently very dissimilar on unobservable variables since they are making choices that are different, that could have a substantial impact on their lives, and that are somewhat costly to reverse. Making a school decision is not one in which there is naturally a lot of randomness (as there might be in a decision, like choosing whether to be light or dark in the game of checkers, where most people would be willing to flip a coin). Thus, one must assume that the students differ on some characteristic(s) that affects their expected outcomes in charter schools or traditional public schools. That is, selection bias must exist.

Another way to understand the frailties of matching is to consider a researcher who attempts to evaluate charter schools by finding pairs of genetically identical twins in which the parents have decided to send one twin to a charter school and the other to traditional public school. Since the parents *know* the twins and could not possibly make such a decision without contemplating how the different treatment would affect them, it must be case that the twins differ greatly on unobservable characteristics that determine their school-related outcomes. In fact, by minimizing the inter-student differences on some variables (as the researcher did by focusing on identical twins) while requiring that the matched students differ on whether they choose treatment or no treatment, the researcher is *maximizing* differences on unobservable variables that affect school-related outcomes. In other words, he is maximizing the impact of selection biases on the estimate. This point has been known for some time (see Neumark 1994) but is nevertheless sometimes forgotten by users of matching methods.

In other words, matching is not always useful but is a method that must be used with careful attention to how it interacts with selection. In some scenarios, selection in and out of charter schools may become more random when matching is used. In other scenarios, such as the twins scenario described above, the matching guarantees that whatever separates the twins is *not* random--since parents are in the ideal position to prevent minor random occurrences from generating major differences in their twins' school-related outcomes.

In short, matching methods involve substantial judgement, and the judgements focus on unobserved variables. It should now be evident that simple rules for using these methods are unlikely to apply, since making rules about the relationships between observed and unobserved variables could never be simple. Nevertheless, there are some rules about how to use matching methods for estimating the effect of charter schools on students' outcomes. The CREDO study violates multiple rules.

First, a researcher who wishes to rely on matching methods is obliged, at a minimum, to indicate *what* valid instruments separate the treated and control students, even if obtaining data on the valid instruments is impossible. (The obvious valid instrument is a lottery variable. If the researcher does not use a lottery-based method, he must propose an alternative instrument.) If the research indicates what the instruments are, readers can assess their plausibility on logical, if not empirical, grounds. If the researcher is so unsure of his instruments that he cannot bring himself to describe them, then readers should dismiss the idea that his matching method produces unbiased results. If obtaining data on the valid instrumental variables *is* possible, an instrumental variables method ought to be applied.[2] The CREDO study does not indicate or even hint at what its instruments are.

Second, matching, for a decision as important as choosing a school, ought never to proceed so far that it is likely that families with matched children who make different decisions *know one another* or have their children known by the same decision-maker. For instance, if two apparently identical students who are in the same fourth-grade classroom make different decisions about the school at which they will attend fifth grade, it is likely that their families know one another, certain that the students know one another, and certain that the teacher knows both students. Therefore, it is unlikely that it is only a minor random occurrence that causes them to make difference school choices--the students, the families, and the teacher are all in the position to lessen the impact of such minor occurrences. For instance, if one family were randomly to hear about the charter school and the other family were randomly not to hear about it, this informational difference would likely disappear through communication between the students, families, and teacher. If the informational difference persists, its persistence is a red flag that the students (or their families) are not alike. The CREDO study makes the error that it matches children who *are* likely to have known one another or who would have had the same teacher.

Third, matching can only be based confidently on variables that are truly predetermined, such as race, ethnicity, gender, and age. Matching ought *not* to be based on variables such as program participation that are partially controlled by schools. For instance, students ought not to be matched on whether they participate in a school meals program, services for the disabled, or services for non-native language learners. To see this, suppose that the traditional public schools are more supportive, philosophically, of putting students into disability services and the charter schools are more supportive of keeping students in mainstream classrooms whenever possible. Then, if "disabled" traditional public school students are matched with "disabled" charter school students, the charter school students will systematically be much more disabled. The matching will have exacerbated, not alleviated, the

---

[2] Note that, once a valid instrument is available, it is often complementary with matching methods, which can improve statistical power and reduce the importance of linearity and distributional assumptions. That is, the data can first be reduced to treated and control samples with similar distributional support for the predetermined covariates. The instrumental variables method can then be applied to these data.

differences between the treated and control students. The CREDO study makes the error that it matches children based on variables that are partially controlled by schools: participation in the free or reduced-price lunch program, participation in special education services, participation in English Learner services.

Fourth, while matching to traditional public schools from which charter school students came may seem like a good idea, it is not a good idea for any charter school that enrolls most children as kindergarteners because they do not have a prior school. The prior schools of students who enter above kindergarten ("late entrants") are often not representative of the schools that most of the charter school students would have been attending if they had not entered the charter school. Matching should be made with the schools that charter students would otherwise have attended, *not* just with schools that late entrants came from. The CREDO study makes the error that it matches children to schools based solely on the late entrants.

Summing up, in addition to the serious statistical mistake described in Section III, the CREDO study violates four rules for matching-based studies:
(i) it does not indicate what valid instrumental variables, if any, determine school choices;
(ii) it matches charter school students to traditional public school students whom they were likely to know (or be known by the same teacher);
(iii) it matches on variables that are partially controlled by schools, such as participation in special education services.
(iv) it matches students to the schools of the late entrants, not to all schools that the charter school students would have attended if they had not entered the charter school.

V. Conclusions

The results of the CREDO study cannot be relied upon. This is most obviously because the study contains a serious statistical mistake that generates an automatic negative bias of considerable magnitude. This mistake is a function of straightforward statistics familiar to those who are aware of measurement error. It is not an obscure mistake that researchers could not be expected to understand. In addition, the CREDO study violates several rules for the empirically sound use of matching. It is not possible to say exactly how such rule-breaking affects the estimates, but the rule-breaking should be fixed.

References

Center for Research on Education Outcomes (CREDO). 2009. *Multiple Choice: Charter School Performance in 16 States*. Stanford, CA: CREDO.