Fact vs. Fiction:

An Analysis of Dr. Hoxby's Misrepresentation of CREDO's Research

CREDO

Stanford University

http://credo.stanford.edu

October 7, 2009

Introduction

The memo, "A Serious Statistical Mistake in the CREDO Study of Charter Schools," by Caroline Hoxby, does not provide any basis whatsoever for discounting the reliability of the CREDO study's conclusions. The central element of Dr. Hoxby's critique is a statistical argument that is quite unrelated to the CREDO analysis. The numerical elements of it are misleading in the extreme, even had the supporting logic been correct. Unfortunately, the memo is riddled with serious errors both in the structure of the underlying statistical models and in the derivation of any bias.

The technical aspects of these issues are described below, but the overall description of the problem is straightforward. She considers estimation of models that are very different from those estimated by CREDO and derives conclusions that are completely irrelevant to the CREDO results.

Summary of the Arguments

The central assertion by Dr. Hoxby is that measurement error in prior achievement is different between the charter school students and the control group in the traditional public schools and that this difference will systematically bias the estimated effectiveness of charter schools downward. To arrive at this conclusion, she considers separate estimations for samples stratified by whether a student is in a charter school or not. For each sample, she considers a simple bivariate regression where student achievement growth is determined solely by prior achievement scores and an intercept. The difference in the intercept estimated for the two separate samples of students is then identified as the charter school effect. As is well-known, in a bivariate model with measurement error of the single explanatory variable, the coefficient on the explanatory variable will be biased toward zero – and it is this premise, misapplied, that is the heart of her critique.

The fundamental conclusions rest completely on this structure, but this structure is very different from the CREDO analysis. The CREDO analysis estimates a pooled model with a single coefficient on prior achievement for both charter and traditional public school students, so the entire derivation of Dr. Hoxby that is based on separate coefficients for the two groups is an incorrect representation of the CREDO models. Further, the CREDO analysis estimates multivariate models with more explanatory factors than simply prior achievement. As is well-known in the statistical literature, there is no longer a presumption that the coefficient on prior achievement is biased downward (even in the specific case where it is estimated from the samples stratified by charter school status). Finally, the

calculation of bias in the chosen formulation by Dr. Hoxby is obviously incorrect and cannot support her back-of-the-envelope assertions that the bias is large.

We describe each of these issues below.

Hoxby's Critique of CREDO's Use of Statistics

Dr. Hoxby asserts that a statistical mistake could cause the CREDO estimates of the effects of charter schools to be severely biased downwards. The bulk of Dr. Hoxby's critique aims to prove that the use of Virtual Control Records (VCRs) or "virtual twins" creates a *negatively biased* estimate of the effect of charter school enrollment on student achievement growth. However, these statistical arguments are an artful construction that permit her to make rather exaggerated claims. In addition, these arguments themselves contain serious errors that point her to a very misleading conclusion.

1) The statistical model described by Dr. Hoxby assumes that a separate relationship of achievement growth and prior achievement is estimated for the charter school students and for the Traditional Public School (TPS) students but in fact a single parameter from a pooled relationship is estimated. We did not set out to compare the absolute magnitude of the charter and TPS effects, but rather to estimate a relative comparison of the impact of attending a charter vs. attending a TPS. As a result, her derivations are simply incorrect. Interested readers can see the Technical Appendix for a numerical representation of the difference between the model Dr. Hoxby mistakenly believes we estimated and the model actually estimated by CREDO.

2) A further flaw in the Hoxby critique is that her statistical model evaluates something different than what was done in the CREDO analysis, allowing her to state unequivocally that any bias in the estimated charter school effect is negative. While it may be convenient and more supportive of her case to consider a model with only prior test scores and school type, as in her equations (3) and (4), these are not the models estimated in the CREDO analysis.[1] The CREDO models include other covariates (e.g., race, and free lunch status). The "bias" that Dr. Hoxby estimates is no longer the simple expression that she provides

[1] Hoxby incorrectly asserts that "Without loss of generality, other covariates on the right-hand side of the equations have been omitted for simplicity."

when one considers a larger number of factors. Importantly, even the sign of any potential bias is no longer known without further specification and knowledge of the underlying structure of the data. Again, please see Technical Appendix for the difference between the CREDO model and Dr. Hoxby's inaccurate specification.

Apart from the serious statistical mistakes in Dr. Hoxby's proof outlined above, her argument is further undermined by her assumption that the development of our VCR sample is subject to the Law of Large Numbers, greatly reducing any measurement error, and therefore introduces significant bias. She asserts that the test measurement error in the VCR test scores will differ significantly from that in the charter school student scores because each VCR record averages the scores of several students. [2] Dr. Hoxby is correct that the VCR record is an amalgam of the experiences of multiple students. However, her invocation of the Law of Large Numbers is used to reach unsupported and frankly outrageous conclusions. Here are the facts: the median number of control student matches is 6, and roughly 75% of the students in our sample have 10 or fewer matches. The neatness and clarity of her derivation disappears when it is recognized that, to the extent that measurement error is relevant, it appears in both the achievement of the charter school students and the majority of VCRs. Simply put, 6 is not a "large number", invalidating this entire line of her argument.

It is also important to remember that even if Dr. Hoxby's analysis were correct and we were in fact estimating stratified samples with different measurement errors, the test measurement error must be large enough to lead to substantial biases and the *difference* in measurement error between the charter student and VCR achievement must be large

[2] Test measurement error may arise (and the operative word is *may*) from a number of different things, including faulty measuring instruments (think of a poorly manufactured ruler) or a bad day for the test taker. It can be positive, negative or entirely absent, but we don't know in any given case if the score is higher, lower, or accurate. If there is bias, the score we see is greater or smaller than a student's real abilities. Typical test reliabilities, which provide an indication of measurement errors, are quite high, indicating that the whole issue may not be particularly relevant.

enough to impart substantial differential bias in the estimated relationship. Neither of these assertions is obvious.

3) Dr. Hoxby herself highlights the arbitrariness of her statistical manipulations when she derives the bias in equation (11). This is incorrect and cannot be used to infer any problems with the CREDO study. Specifically, she asserts, without providing any details, that the "bias" in the estimated overall charter school effect – something that should be a single value – varies by individual student, year, and school ($\zeta A_{ij,t-1}$). Dr. Hoxby seems to have confused measurement error in individual student test scores with the resulting bias in the estimators. Of course, it does not make any sense to think about 1,700,000 separate values of any bias in the CREDO study.

4) Her attempt to indicate the magnitude of bias is arbitrary and exaggerated; further, it does not reflect other analyses in the literature. Specifically, in her formulation of equation (4), μ_1 would be typically interpreted as the depreciation rate of knowledge, i.e., how much achievement falls off over the summer and from year to year.[3] The rough estimate that Dr. Hoxby uses to quantify any bias assumes that there is no depreciation of knowledge and that any nonzero estimate of μ_1 found in prior studies reflects test measurement error. Her assumption and interpretation is a capricious exaggeration, and it is entirely inconsistent with the common interpretation of such models. For her to extrapolate that the measurement error must be at least 0.2, and that this effect will compound to at least a bias of 0.4 is therefore unsubstantiated. Moreover, when the importance of test measurement error in models with lagged achievement has been investigated, its effect on the estimation has been found to be small.[4]

[3] Note that in equation (4), μ_1 would be expected to be zero with no depreciation. If the left hand side is just $A_{ij,t}$ and not $A_{ij,t} - A_{ij,t-1}$, the expectation for μ_1 without depreciation of knowledge would be 1.

[4] Hanushek, Eric A. 1992. "The trade-off between child quantity and quality." *Journal of Political Economy* 100,no.1 (February):84-117.

Hoxby's Critique of Matching Methods

In addition to the claims made about the effects of test measurement error, Dr. Hoxby also critiqued the use of matching methods in general and the CREDO approach in particular. These arguments are even less defensible than her improper use of statistics, but for the sake of a complete response, they are addressed below, organized around her four pronouncements.

1. "We fail to include (or mention) suitable instruments that determine school choice."

It is true that an instrument may be useful to ensure random charter school placement, but this is not necessary from an empirical standpoint. There will only be selection bias to the extent that unobserved differences between charter and TPS students affect their achievement gains. Dr. Hoxby also assumes that the decision function for each parent includes a consideration of their child's academic learning gains at a charter vs. a TPS, even among siblings. To the extent that charters are chosen for other reasons (e.g. safety, discipline, etc.) or for parents whose decision functions did not include as an input a considered comparison between their child's academic growth at their local TPS and charter, by definition this "unobserved variation" will lead to no selection bias as it relates to academic achievement.

2. "We match charter school students to TPS students they are likely to have known, or who will have been known by the same teacher."

This is a theoretically possible idea that is not at all a plausible cause of significant bias in our aggregate results. First of all, since we pool all the TPS students from all the schools that lose students to a given charter, it is likely that the students in a VCR come from different schools. Even if the student who is in a charter school comes from the same school as their match, the odds still likely favor VCR students from the same school having different teachers. Thus, it is improbable that a charter school's children actually knew the students who are included in their VCR match, or that any given teacher knew the charter student, or the students included in the VCR. But even if the scenario were true, for this to be a cause of bias other conditions also must be met:

1. The students, or their families, or their teachers must know each other, and well enough to discuss various forms of schooling.

2. This information is passed along from student to family to teacher or among them in a free flow of information such that all students, parents, and teachers are equally knowledgeable with respect to their school choice options.
3. The knowledge must induce systematic selection into charter schools.

Each of these situations must be true for Dr. Hoxby's point to hold, that by selecting students from the same schools with the same information set available to them, the persistence of an information gap is itself evidence that there are achievement-related differences among the students or families. The probability that all these conditions would be met for the majority of students, as Hoxby's point asserts, is infinitesimally small.

3. "We match on variables partially controlled by schools, such as participation in special education services."

If participation in programs partially controlled by schools should not be used as a matching mechanism, should it also not be used as a control variable? And if so, does Dr. Hoxby believe that every study which uses eligibility for free/reduced lunches as a proxy for poverty, special education services for disability, and ESL participation for ELL status are in the wrong?

Apart from this, she fails to fully develop her point, likely because it would weaken it significantly were she to do so. The example she gives of potential bias is one in which a charter is less philosophically inclined to label students as special education, which is true in some charters. However, it is entirely possible that for special education, ELL or free/reduced lunch classification, any given TPS could be more or less philosophically inclined to label children than the charter(s) it feeds to. Bias will exist in our estimation only to the extent that there is a) a difference in the process by which students are classified, b) this variance is significant enough to manifest itself in test score growth measurement, and c) these variations are large enough to be distinguished from the natural variation in "achievement impact" among the demographic classifications. For example, since we do not break out the various special education codes, there will be significant variation in the extent to which different determinations may affect test score growth. Any variation caused by philosophical differences must be systematic across the sample and large enough to distinguish itself among this noise.

4. "We match students to the schools of late entrants, not the schools that charter students would have attended."

Dr. Hoxby seems to think that most communities have vast numbers of infinitely accessible local schools and that parents who start their children in a charter would have selected completely different traditional public schools than parents who move their children later on. In all but the largest communities, the pool of traditional public schools is small to begin with and each loses students to the charter. The theoretical collection of schools Dr. Hoxby cites and the actual schools that draw transfer students are in most cases one and the same. This renders her point invalid.

Discussion

It is useful to have an exchange in order to permit cataloging of the pluses and minuses of different analytical strategies designed to uncover the impact of charter schools. One clear truth is that there is no "gold-standard" methodology that is both independent of the question being addressed and beyond reproach. If one wishes to investigate the effectiveness of a given popular school compared to alternative schools that the student might have attended, the use of school lotteries has strong appeal, particularly if one can presume that the lotteries are truly random, that there are sufficient numbers of students involved, and that students can be followed outside of the charter school. But, if one wishes to investigate the effectiveness of a charter school that does not have a lottery – due to legal requirements, lack of sufficient parental demand, or any other factor – other approaches must be used.

The matching methods of CREDO do rely upon obtaining control students that differ from charter school students in random ways. The approach taken was designed to do this as accurately as possible – and it is difficult to verify how accurately this was done. As discussed, while it is always important to consider the potential for measurement error to introduce bias in analyses of achievement, the focus on it to explain the CREDO results is neither logically nor numerically supported by her arguments.

Moreover, one must consider the implications of student selection in the CREDO analysis. If charters had no advantage or disadvantage over traditional public schools, the resulting estimates of school effects in the CREDO study would suggest that, on average, less motivated students attended

charters (after consideration of prior achievement, race, income, etc.). This may in fact be the case. Or they might find other non-achievement aspects of charters worth any loss in quality along the achievement dimension. These are issues to investigate further.

Conclusion

Our hope is that we can get beyond the distractions of this technical discussion and return to the larger policy issues.

It should be noted that this peer-reviewed CREDO study found that charter school performance varied considerably. Some communities and states have gotten the policy right and are able to demonstrate positive charter school results. The recent results for New York City (NYC) indicate that it, too, has a focused and effective charter policy. Along with states like Arkansas, Louisiana and Missouri, these results offer a glimpse into the realm of what is possible when the barriers to entry are removed, the authorizers are high functioning and poor quality schools are removed from the field.

On the other hand, the CREDO study shows that nationally there are more poorly performing charter schools than good ones, leading to obvious and important questions about why some charters in some communities do better than schools in other environments.

It is quite important to understand what determines the differences between the strongly favorable results for NYC and other high-quality states and the more mixed results for a typical charter school elsewhere in the nation.

The critique of the CREDO study by Dr. Hoxby provides no useful guidance for research and policy judgments. It implicitly takes the position that the different results of the studies were purely methodological and not the result of differences in the operation of charters. This is an unfortunate distraction from what could be a substantive discussion of serious policy issues related to the expansion of choice programs under charter schools.

The stakes for charter schools are incredibly high, and so too the research on their effectiveness. Discussions of designs and methods will and should continue in an effort to provide stakeholders and constituencies with high quality analyses. We remain committed to the pursuit of fair and balanced tests

of charter school effectiveness, and are open to feedback on how to improve our work. Dr. Hoxby's research and ours have shown the strong potential of well-run charters operating in a favorable policy and authorizer environment, even if those narrow results do not reflect the full population of charter schools. Our research, given its national scope, asks a broader performance question and holds the charter movement at large to account. We seek to hold accountable those schools that are unwilling or unable to improve and thus put the entire charter movement at risk. In this sense, there is no reason to presume that CREDO's research and that of Dr. Hoxby are in disagreement; in many respects they are complementary. We at CREDO continue to aim to find what works for students, parents and policymakers. CREDO welcomes continued dialogue on these important matters with Dr. Hoxby as well as the rest of the national research community.

Technical Appendix

Equations 1 and 2 below represent the stratified achievement models Dr. Hoxby mistakenly believes CREDO estimated in our study.

$$A_{ijt} - A_{ijt-1} = \mu_0^{TPS} + \mu_1^{TPS} A_{ijt-1} + \varepsilon_{ijt} \quad (1)$$

$$A_{ijt} - A_{ijt-1} = \mu_0^{Charter} + \mu_1^{Charter} A_{ijt-1} + \varepsilon_{ijt} \quad (2)$$

Equation 3 below represents the actual model of achievement the CREDO study estimated.

$$A_{ijt} - A_{ijt-1} = \mu_0 + \mu_0^{Charter} + \mu_1 A_{ijt-1} + \varepsilon_{ijt} \quad (3)$$

In Dr. Hoxby's incorrect specification, there are two separate estimates for data sets stratified for TPS and charter students, with two corresponding intercepts, μ_0^{TPS} and $\mu_0^{Charter}$. The bias that Dr. Hoxby identifies in the estimate of the charter school effect comes from her conclusion that $\mu_1^{Charter}$ is biased and differs from μ_1^{TPS}. CREDO actually estimated a common coefficient μ_1 for the entire sample, as shown in equation 3. It simultaneously directly estimated the charter school effect, $\mu_0^{Charter}$, based on the entire sample and the model in equation 3. Even in the simple model of equation 3, if measurement error in prior achievement (A_{ijt-1}) does affect the estimate of μ_1, the implied impact on $\mu_0^{Charter}$ in equation 3 is not found by the expressions included in Dr. Hoxby's memo, and it is no longer a simple function of the variance in the measurement error in A_{ijt-1}.

Equation 4 below shows the actual multivariate model CREDO estimated. Contrast this with the simple, and incorrect, bivariate model Dr. Hoxby uses to represent the CREDO model in equations 1 and 2. This is the simplest model that CREDO estimates.

$$A_{ijt} - A_{ijt-1} = \mu_0 + \mu_0^{Charter} + \mu_1 A_{ijt-1} + \mu_2 (race) + \mu_3 (ell) + \mu_4 (special\ ed) + \mu_5 (grade) + \mu_6 (year) + \mu_7 (retained) + \mu_8 (Katrina) + \mu_\zeta + \varepsilon_{ijt} \quad (4)$$

where μ_ζ represents a fixed effect for the state where each school in the sample is located.

In the multivariate model estimated by CREDO it is simply not the case that μ_1 is always biased toward zero in the manner Dr. Hoxby presumes, as any bias depends on expressions which include the variances and covariances of A_{ijt-1} with each of the other independent variables in the model.