



School of Education, University of Colorado Boulder
Boulder, CO 80309-0249
Telephone: 802-383-0058

NEPC@colorado.edu
<http://nepc.colorado.edu>

REJOINDER OF ANDREW MAUL TO RESPONSE FROM CREDO

SEPTEMBER 28, 2015

Background

On three occasions over the past two years, I reviewed reports on the effectiveness of charter schools published by the Center for Research on Education Outcomes (CREDO). The reviews, published by the National Education Policy Center, were mixed, noting strong elements of the research design while also raising concerns about the methods used and the report authors' characterizations of the implications of their findings. These reviews were of the 2013 CREDO national study, the 2013 study of Michigan charters, and the 2015 national study of charters in urban regions.

In June of 2015, CREDO released a statement responding to some of the critiques I raised in my review of the most recent of these—the 2015 CREDO study of the performance of charter schools in urban regions—and to a separate review by Andrea Gabor, the Bloomberg Professor of Business Journalism at Baruch College/CUNY.

My response to the CREDO statement follows. First, however, I want to thank the scholars at CREDO for taking the time to read and respond to the concerns about their work I raised in my reviews. At the end of the day, our priorities are aligned: we all agree on the critical importance of improving the quality of education for all students, and on the importance of rigorous research in pursuit of this goal. Although we may have different perspectives on a number of specific points, I hope that this kind of public

exchange has the potential to improve the overall quality of educational policy-relevant research.

Eight Important Concerns about CREDO's Charter School Studies

1. The Nature of the Comparison between Charter and Traditional Public Schools in the CREDO Studies is Not Clear

In my review, I expressed concern about the lack of clarity and transparency regarding the CREDO study's selection of virtual twins from "feeder schools" (i.e., the traditional public schools from which students in the study left to enroll in charter schools). The CREDO response to this concern states: "Dr. Maul and others may wish for a head-to-head sector-to-sector study, but that is not the study we intended to perform. Our hypothesis is a student-level question: how well are charter students progressing?" The CREDO response then goes on to provide more detail about the reasoning behind their choice to focus only on comparing charter students to matched students in feeder schools, noting that "we have never asserted that the VCRs [Virtual Control Records] are representative of the full range of traditional public school students because we know they are not."

I appreciate the clarification of the rationales for this choice of comparison. However, it seems to be based on a misreading of my initial concern. I never expressed any "wish for a head-to-head sector-to-sector study" – in fact, I explicitly stated that the exclusive use of feeder schools was "arguably a reasonable choice" (p. 5). My concern was and is that this detail appears in neither the main report nor the Technical Appendix, and therefore the CREDO report invites the misinterpretation that the study provides a comparison between charter schools and Traditional Public Schools (TPSs) in general.

I would note that this misinterpretation seems to be pervasive in the media attention surrounding the CREDO report. One commonly finds claims such as that the CREDO report provides evidence that "students in charter schools learned significantly more than their peers attending traditional public schools." This is not the sort of statement that can be made, even in principle, on the basis of the CREDO study. Thus my concern stands.

2. The Matching Variables Used in CREDO's Studies May Not Be Sufficient to Support Causal Conclusions

In the context of expressing my concerns regarding the VCR (Virtual Control Record) matching technique, I noted that a leap of faith is required to believe that the seven matching variables employed in the study are truly sufficient to account for all

meaningful differences between charter students and their feeder Traditional Public School (TPS) peers. As an illustration, I noted that “one must believe, for example, that the dichotomous ‘eligibility for subsidized meals’ variable (along with the other demographic variables) is sufficient to control for all meaningful socioeconomic differences in students” (p. 4).

The authors of the CREDO response “agree to a point” with my concern, but note that “better measures of socio-economic status have not gained any traction in education policy circles,” and conclude that “FRL [Free and Reduced Price Lunch] eligibility ... is the single most appropriate indicator of economic status at this time.”

In response, I can only say that I sympathize with the authors’ frustration over FRL eligibility being a “blunt instrument and not as sensitive a measure as one might desire.” This is indeed a perennial frustration for many policy researchers. Unfortunately, it does nothing to alleviate my concern, as it remains the case that, in order to defend the sorts of causal statements made in the CREDO report, one must believe that FRL eligibility (in combination with the other demographic variables) is sufficient to control for all meaningful socioeconomic differences in students.

To be clear, my concern relates to the authors’ decision to include causal statements about charter school effectiveness, not about the inclusion of FRL eligibility or the other variables. Thus my concern stands.

3. Some Lower-Performing Charter Students are Systematically Excluded from the CREDO Studies

In my reviews, I noted a general concern about the CREDO study’s use of a “home-grown” matching technique (CREDO’s VCR method) instead of the more established and well-understood technique of Propensity Score Matching (PSM). I noted several specific reasons for this concern, one of which was that:

the VCR technique found a match for only “greater than 80%” of charter students (Technical Appendix, p. 8), meaning that close to 20% of charter students were excluded from the study. In [an earlier CREDO report] it was indicated that the excluded students had average scores 0.43 standard deviations lower than the average of the included students, introducing the potential for bias due to systematic exclusion of lower-performing students (p. 3).

In response to my concerns, CREDO’s June 15 statement claims that “the parameters of propensity matching have been explored in full against the VCR method by an uninterested research group and was [sic] found to be inferior. See Gleason et al. [2010]”

I am a bit baffled by this response. The cited study does not contain even a single mention of propensity-based techniques, let alone a comparison of PSM and the VCR technique. Thus my concern stands.

4. CREDO's Reasons for the Systematic Exclusion of Lower-scoring Charter Students Do Not Address the Potential for Bias Arising from the Exclusion

In response to my concern regarding unmatched charter students having average scores 0.43 standard deviations lower than the average of the included students, as described in the previous section, the authors of the CREDO response claim that “Dr. Maul misinterprets the implication of unmatched charter students. The charter students who were unmatched were excluded because there were no matches for them in the traditional public schools in the feeder pool ... not because of their prior academic performance ... but because they present unusual combinations of the remaining factors.”

This response similarly baffles me, as I cannot see how my concern is affected by the specific reasons for the exclusion of these students. It remains the case that a substantial number of apparently lower-performing charter students were systematically excluded from the analysis (and that this likely would not have happened had PSM been used instead of the VCR method). This introduces the possibility of bias into the comparison of charter students and feeder TPS students. Thus my concern stands.

5. The “Days of Learning” Metric Used in the CREDO Studies is Problematic

In response to my concerns about the use of “days of learning” as a translation for effect sizes, the authors of the CREDO report reference “the work of Hanushek, Peterson and Woessman” and state that that results are “not presented in scientific notation [sic] since that is not the style of report we prefer to release.”

I can sympathize with the desire to make the concepts of standard deviations and effect sizes accessible to as wide a readership as possible; however, oversimplification is also a concern. While these concepts are not terribly complicated, neither are they so simple that it can be assumed out of hand that a simple translation into a time metric will be sufficient to communicate their full meaning to a wide, nontechnical audience. For this reason, when communicating to such wide audiences, it would clearer for researchers to give several ways to interpret effect sizes, such as the percentage-of-variance-explained metric (discussed further in Point #8, below).

As I discussed in my original review of the CREDO report, the conversion of effect sizes into “days of learning” requires several heavy assumptions, many of which are still hotly contested in the literature on vertical scaling. Moreover, when average year-to-year changes in test scores are fairly small, nearly any difference between groups will seem large when expressed in this metric. For example, in Hattie’s (2009) famous meta-meta-analysis of studies of educational interventions, it was reported that the average effect of an educational intervention was .40 standard deviations. CREDO’s conversion table only goes up to an effect size of .30, but by extrapolation an effect size of .40 would be equivalent to 288 days of learning – more than ten times the size of many of the effect sizes reported in the CREDO studies.

Many educational interventions yield effect sizes far larger than this, of course; for example, the use of “Piagetian programs” is associated with an effect size of 1.28, or 922 days of learning (and, notably, this finding is still the subject of considerable controversy). On a more whimsical note, a recent study (Chaddock-Heyman et al., 2015) found that the difference between being in the upper and lower thirds of cardiovascular fitness was associated with a .59 standard deviation difference in math scores (425 days of learning).

Referring to effect sizes of .05 standard deviations as evidence of the “smashing success” of charter schools is hyperbolic. Thus my concern stands.

6. The CREDO Studies Fail to Provide Sufficient Information about the Criteria for the Selection of Urban Regions Included in the Studies

In response to my concern about insufficient detail being given regarding the criteria for selection of urban regions to be included in the study, the authors of the CREDO report express confusion and state that “the Technical Appendix details the methods used to select the 41 regions under study.” However, the Technical Appendix only lists the criteria that were considered (total city population, total population of charter students, total number of urban charter students, size of the primary school district(s), and charter market share). It does not discuss *how* these criteria were utilized in the selection of the 42 (not 41) regions included in the study. This falls well short of providing enough information to permit replication of the study by independent researchers, which is the usual standard for academic publications. Thus my concern stands.

7. The CREDO Studies Lack an Appropriate Correction for Multiple Significance Tests

In response to my concern about the potential for an inflated familywise Type I error rate due to the lack of a correction (e.g., a Bonferroni adjustment) for multiple significance tests, the authors of the CREDO report state that “this correction does not apply to CREDO’s studies because we are not saying there is an effect if ANY one comparison is significant. We instead report the number and degree of significant findings.” However, to the extent to which individual coefficients are interpreted, including results for individual states and individual urban regions, the concern does apply – the chance of a false positive occurring somewhere along the line is far greater than 5%. Thus my concern, while minor in this case, stands.

8. The CREDO Studies Have Trivial Effect Sizes

My biggest concern with the CREDO studies is that reported effect sizes are so small in magnitude that they may be regarded as trivial. As one way of expressing this, I noted that “the overall effect sizes reported are 0.039 standard deviations for reading tests and 0.055 standard deviations for math tests. If they were correct, these numbers could be interpreted as stating that well less than a tenth of one percent of the variation in test scores can be attributed to whether a student is in a charter school or a ‘feeder’ traditional public school.” In response, the authors of the CREDO study make two comments. The first is that “CREDO looks at academic growth, not academic status.” The second was that “multiple individuals both at CREDO and elsewhere have tried to recreate the statistic that Maul reports and have been unable to do so.”

With respect to the first comment, the method CREDO uses for modeling growth is a regression of present academic status on a treatment indicator variable, as well as prior academic status and other control variables. This is a standard method commonly used throughout the literature. CREDO additionally subtracts out prior academic status from the dependent variable and refers to this variable as “growth,” but this subtraction does not influence the estimation of the treatment coefficient and is therefore irrelevant.

With respect to the second comment, as briefly discussed in section 5 above, there are a number of methods for expressing effect sizes. When the mean values of a continuous variable (e.g., test scores) are compared between two groups (while controlling for prior test scores and other variables), the most common way of expressing effect sizes is in terms of standardized mean differences, or Cohen’s *d*, which is how they are expressed in the CREDO reports (that is, prior to the conversion to “days of learning”). Another common mode of expression is in terms of the percentage of variance in the outcome variable (again, test scores) that can be uniquely explained by variance in the predictor

variable (whether a student comes from a charter or “feeder” TPS), net of all other variables; this is sometimes referred to as R-squared (R^2) or eta-squared (η^2).

The conversion from Cohen’s d to eta-squared is given by Cohen (1988) as:

$$\eta^2 = \frac{f^2}{(1 + f^2)}$$

where $f = 2d$.

The most recent CREDO report estimates the overall effect sizes (Cohen’s d) to be 0.039 standard deviations for reading tests and 0.055 standard deviations for math tests. Using the above equations, the value of eta-squared is .0008 (.08%) for mathematics and .0004 (.04%) for reading. That is, under one tenth of one percent of the variance in final test scores can be uniquely explained by whether a student is enrolled in a charter or a “feeder” TPS school.

Thus my concern stands.

The Bottom Line: The Findings of the CREDO Studies Cannot be Regarded as Compelling Evidence of the Greater Effectiveness of Charter Schools Compared with Traditional Public Schools

I appreciate the willingness of CREDO researchers to engage with my reviews. This stated, I do not find that the CREDO response adequately addresses any of the concerns that I have raised, and my final assessment remains unaltered: “the findings of [the CREDO reports] cannot be regarded as compelling evidence of the greater effectiveness of charter schools compared with traditional public schools, either overall or specifically within urban districts.”

References

Chaddock-Heyman L, Erickson KI, Kienzler C, King M, Pontifex MB, Raine LB, et al. (2015) The Role of Aerobic Fitness in Cortical Thickness and Mathematics Achievement in Preadolescent Children. *PLoS ONE 10(8): e0134115*.
doi:10.1371/journal.pone.0134115

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd ed.* Hillsdale: Earlbaum.

Hattie, J. (2013). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* Routledge.

Maul, A. (2015). *Review of "Urban Charter School Study 2015."* Boulder, CO: National Education Policy Center. Available online at <http://nepc.colorado.edu/thinktank/review-urban-charter-school>.

Maul, A. & McClelland, A. (2013). *Review of "National Charter School Study 2013."* Boulder, CO: National Education Policy Center. Available online at <http://nepc.colorado.edu/thinktank/review-credo-2013>.

Maul, A. (2013). *Review of "Charter School Performance in Michigan."* Boulder, CO: National Education Policy Center. Available online at <http://nepc.colorado.edu/thinktank/review-charter-performance-michigan>.