NATIONAL EDUCATION POLICY CENTER

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

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

Response of Sara Goldrick-Rab to the Authors' Reply

September 17, 2012[1]

I appreciate that Matt Chingos and Paul Peterson, unlike some voucher supporters, are willing to engage with my critique of their study on the actual substance (see their response in *Education Next)*.

Their responses, however, seem to suggest that they missed several key points in my review, and so here I provide further elaboration. As always, I encourage all who want to engage in academic discourse to read in full the points on both sides before jumping into the fray.

First, to be quite clear and direct, I had – and still have – two primary objections.

First, I am concerned about the accuracy of the estimated impact of vouchers for African American students, and the manner in which it was presented as a conclusive finding, despite social scientific convention:

- The authors' estimation of the treatment impact is more imprecise than as stated in the report, because of (a) non-random measurement error and (b) meaningful imbalance in the pre-treatment characteristics of the treatment and control groups. Since the authors haven't made the numbers available, we cannot know whether the

[1] I'm indebted to Howard Bloom, Felix Elwert, Steve Raudenbush, and Chris Taber for useful conversation on these points.

estimated impact for African Americans would be statistically significant if these concerns were taken into account. But it is very possible that they would not.

- When the average treatment impact estimate is null, when only one subgroup has a statistically significant impact, *and when the construction of subgroups was not performed in the original random assignment,* it is unconventional in experimental research to describe results as conclusive and recommend policy changes based on them (as did the authors). The risk of a Type 1 error (false positives) is too high for this to be advisable, and thus these sorts of results are best described as "exploratory" and worthy of further research.

Second, I take issue with how the results are displayed in the report, and the authors' apparent unwillingness to unveil their full findings so that readers can examine them.

Their response offers no additional data and instead simply asserts that they are correct. This is not helpful to the research community or to the policymakers and practitioners who need reliable information to help students. If Chingos and Peterson did not feel that it was appropriate to present the requested data and information in *Education Next,* fair enough – but they should do so elsewhere. Why haven't the authors made this key information available on, for instance, the *Education Next* website?

Below, I elaborate on both types of concerns. Let's start with the methodological issues.

Validity of the Impact Estimates

As summarized above, I have two substantive concerns with the validity of the impact estimate—for African American students in particular, but also in general for all groups: (a) threats from measurement error, and (b) threats from baseline non-equivalence.

Measurement Error

My critique raised detailed concerns with measurement of the dependent variable (college enrollment), and Chingos and Peterson responded by claiming that all measurement error already appears in the standard errors. This is not true. The authors mistakenly assume that the measurement error in the dependent variable is random.

In actuality, the measurement error I am describing is differential error. The National Student Clearinghouse (NSC) indicator of college enrollment relies on the accurate identification of a matched student record in its database—if none is found, college enrollment is recorded as zero. Differential measurement error may stem from incomplete coverage of colleges and universities (this is the one problem that the authors note) as well as from uneven coverage of colleges and universities by *sector,* which is demonstrably correlated with race/ethnicity. Economist Susan Dynarski's forthcoming analysis finds that the NSC has less coverage at private and for-profit institutions, where African Americans are overrepresented and where Chingos and Peterson report their only significant estimated impacts on college choice (see Table 6). In other words, the report's authors lean heavily on estimates that are at risk for differential error that is correlated with the outcome.[2] This is problematic since statisticians have demonstrated that "an analysis ignoring differential measurement error may considerably overestimate the causal effects."

I suggest that the authors to obtain an alternative data source with which to measure college attendance and see if the results replicate. That is what researchers have typically done in studies using NSC data. If this is not possible, then they should undertake sensitivity analyses such as those recommended by Imai and Yamamoto:

> Imai, K. and Yamamoto, T. (2010). "Causal inference with differential measurement error: Nonparametric identification and sensitivity analysis." *American Journal of Political Science*. 54(2): 543-560.

Baseline Non-Equivalence

My critique also points out that there is cause for concern regarding internal validity of the estimates due to some baseline non-equivalence, as illustrated in the report's own Table 1. In their response to me, Chingos and Peterson state that they are completely certain that the treatment and control groups are comparable pre-treatment, based on the results of a test for joint significance of the equivalence of groups based on observable characteristics. They are over-confident. While a test for joint significance using a set of observables might indicate successful randomization of groups, if a key variable is imbalanced that is strongly predictive of the outcome (which at least one is: parental education) there may be cause for concern. We cannot be confident the African American subgroups were equivalent at baseline.

The test for joint significance says nothing about the relative importance of those observable characteristics for the outcomes of interest, nor does it say anything about the equivalence of unobservable characteristics. If the authors think they are right and that I am raising something that does not affect the estimated treatment impact, they should prove this by undertaking sensitivity testing of the type that I pointed to in my original critique:

---

[2] As addressed in my original critique, there may also be differential error due to problems with the matching algorithm. Some analyses, such as this one, and this one indicate that matching errors are correlated with race as well. These differences may attenuate effects, or overstate them—it is hard to know. It would also be useful for the authors to report whether there is differential missingness (by treatment status) due to FERPA blocks on NSC records, as this study does.

Altonji, J., Elder, T., & Taber, C. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy*, *113*(1), 151-184.

Are these concerns likely to matter? Examine the report's Table 3 and notice that the impact estimate for African Americans is just barely significant with covariate adjustments, and is not significant at the customary 5% level once adjustments are included. Given this, it is surprising that the authors do not even present covariate-adjusted estimates in Table 6, which displays results that they spend a fair bit of time emphasizing—the impacts on private and selective college attendance. Is this because those results are also not significant once baseline differences are accounted for? Readers are given very limited information and therefore cannot know.

Interpreting the impact estimates for subgroups

My critique questions the report's emphasis on findings for one subgroup. Had Chingos and Peterson framed the finding for African Americans as an encouraging, exploratory hypothesis deserving of further testing, I would not have been alarmed by the report. But the study's results absolutely do not merit headlines such as "Vouchers promote college attendance for African Americans."

In response, Chingos and Peterson incorrectly assert that it is appropriate to emphasize the findings in the one and only significant subgroup and to frame it as causal and conclusive (appropriate for policy action). I strongly disagree—this is not the convention, given that there are no significant estimates either for the overall sample or for other subgroups. These assertions are also inappropriate given the study design. When random assignment is not blocked/stratified on the subgroups *a priori*—when the construction of subgroups was not performed in the original random assignment—it is *not* common practice in experimental research to put such stock in subgroups, and is not recommended by statisticians who do this work frequently. That is because it heightens the possibility for over-interpreting findings caused by random error (a Type 1 error). Statisticians Bloom and Michalopoulos are clear on this point, writing:

> *Other things being equal, findings for a specific subgroup should not be highlighted unless they differ statistically significantly from those for other sample members. If subgroup differences are not statistically significant, findings for the full study sample usually should be emphasized instead of those for the subgroup. (p. 5)*

The evidence in the Brookings paper represents Bloom and Michalopoulos' "Case 3," which arises when impact estimates are statistically significant for only one subgroup. They write:

> *In this case, we recommend that, other things being equal, results for each subgroup should be considered exploratory. The rationale for this recommendation is as follows. First, the estimated effect is not statistically significant for the full study sample. Hence, the most precise estimate that exists does not provide evidence that the intervention is effective. Second, the estimated effects for the two subgroups are not statistically significantly different from each other. Hence, there is not strong evidence that the statistically significant result for one subgroup is in fact different from the non-statistically significant result for the other subgroup. Consequently, the best information that exists for both subgroups is the full-sample finding.* (p. 12)

The Bloom and Michalopoulos approach is widely embraced by researchers, including those writing about randomized trials of educational interventions. Here are recent examples of researchers who appropriately describe similar findings as exploratory rather than confirmatory.

> Edmunds, J., et al. (2012). "Expanding the Start of the College Pipeline: Ninth-Grade Findings From an Experimental Study of the Impact of the Early College High School Model." *Journal of Research on Educational Effectiveness.*

> Hill, C., Gormley, W., & Adelstein, S. (2012). *Do Short-Term Effects of a Strong Preschool Program Persist?* Georgetown University.

Incomplete Display of Research Findings

While it may have always been the authors' intentions to examine subgroup impacts, it is strange that they only display information for selected subgroups. My critique asks whether there is a negative impact estimated for white and Asian students, and regardless, why are results for those students not displayed? Chingos and Peterson respond only with the assertion that there is an imbalance for that group, and thus the results should not be interpreted. Ok, but why not at least present them?

On first glance, readers may wonder why I suspected a negative impact, given that the effect for African Americans is positive, and the point estimate for Hispanics is also positive (though barely so). If the confidence intervals were wide enough (which they are), then pooling the sample could lead to an overall insignificant effect even if the Hispanic point estimate is positive. However, the point estimate for African Americans is large (0.71) and statistically significant, and the point estimate for Hispanics is small (0.017)—but the overall full sample estimate is smaller still (0.006). Averaging two larger point estimates together, even with a lot of error, should not result an estimate smaller than either of the two subgroup estimates.

It appears that my hypothesis was correct, since in direct correspondence with me the authors admitted that there is a negative estimate for the omitted students, but they claimed that the group is small (about 100) and imbalanced at baseline. There is something strange here, since based on simple math from the tables the report does include, the group of omitted students is in fact three times the size they claim (over 300), and they fail to demonstrate the imbalance in that reasonably sized group in Table 1. I cannot explain this discrepancy.

Oddly enough, their response to me (the one published in *Education Next* online) never mentions the negative impact estimate and does not quibble with my point about the differences in reported versus calculated sample sizes. As stated above, I believe that Chingos and Peterson should have simply displayed the full set of results and allowed readers to judge for themselves. In my view, it is surprising and obstructive that any subgroup impacts were not reported, to say nothing of negative subgroup impacts in an analysis that is almost exclusively focused on subgroup impacts.

Researchers know that estimated impacts based on small samples require cautious interpretation; it's not the place of authors in these circumstances to determine which results are and are not important enough to present.

But let's just say it was an oversight. I would encourage them to do so now, if for no other reason than to assuage any concerns that they are "hiding data."