WORKING PAPER #4
PRINCETON UNIVERSITY
EDUCATION RESEARCH SECTION
OCTOBER 2003
http://www.ers.princeton.edu

Rejoinder to Peterson and Howell

Alan B. Krueger
Princeton University

Pei Zhu
Princeton University

October 2003

Rejoinder to Peterson and Howell by Alan B. Krueger and Pei Zhu, Princeton University[1]

Our original paper uncovered the following six problems with the estimates in Howell and Peterson (2002, hereafter HP):

1) They incorrectly defined the strata controls that delineated the randomization process.

2) They used incorrect sample follow-up weights.

3) They threw out students with missing baseline scores, who comprised more than 40 percent of their sample.

4) Estimates for African American students were sensitive to the sample used in the regression. For a sample that includes students without initial test scores, the voucher effect is statistically insignificant if baseline controls are omitted, or if the controls were selected to minimize the standard error of the estimate.

5) They did not collect data on the race of students; instead, they used a nonstandard racial/ethnic classification based on mothers' race/ethnicity that is inconsistent with U.S. government guidelines.

6) The Two-Stage Least Squares (2SLS) model they estimate is asymptotically biased upwards because of partial compliers and measurement errors in private school enrollment.

In their response, Peterson and Howell (hereafter PH) correct their mistakes regarding the first two points, and use the appropriate strata controls and weights, which reduce their estimate by 10 percent in year 3.[2] For some reason, PH say that they still believe their 2002 book, which uses the faulty data, provides "the best" estimates. Nevertheless, we can take the subject of weights and strata controls off the list of disputed issues.

They reject our remaining four points. Below we explain why their responses are unconvincing. From the outset, we should also make clear that we have not been able to replicate PH's results or their precise sample. Moreover, their results and sample do not match those in Myers and Mayer (2003) -- although ours do. Sometimes PH's sample is bigger, sometimes

---

[1] We thank David Card, Jesse Rothstein, Cecilia Rouse, Marta Tienda for helpful discussions. Krueger gratefully acknowledges financial support from the Carnegie Scholars Program.

[2] They do this by cryptically reporting that MPR revised the weights and strata controls "after certifying an original set." Myers and Mayer (2003) report that MPR revised their weights and strata controls in response to our article. MPR also denies ever having "certified" weights and strata controls for PH.

smaller. To try to resolve this puzzle, we and the editors asked PH to provide the case id's for their samples, but they declined. The discrepancies are greater when covariates are included. Unfortunately, we found it impossible to determine why their results could not be replicated.

Baseline Scores

Greene and Peterson (1996) have previously made the following methodological point, with which we are in agreement:

> Let us repeat: Analysis of randomized experimental data does not require controls for background characteristics or test scores. Such controls are necessary only when one doubts that the experimental data are truly random. The fact that the estimated effects remain essentially the same when these factors are controlled lends further weight to the conclusion that … no critical departures from randomness seem to have occurred.

For three reasons, however, PH have since argued that it is essential to limit the sample to those with baseline scores in the New York experiment. In PH (2002), they argued that there was a difference (p=.105) in baseline scores between the treatments and controls for those who had such scores. Because they could not tell if there was a similar difference for those with missing scores, they argued that this baseline difference made it essential to control for baseline scores. We pointed out to them, however, that their claimed difference was a mistake -- the actual p-value was 0.82 -- and this point was dropped from PH (2003a).[3] Rather than acknowledge that the lack of a T-C difference in baseline scores suggests that there were *not* "critical departures from randomness," however, they have pursued other lines of argument.

PH (2003a) argue that baseline scores are required because "the risk of mismatching was reduced because students put their own names on the baseline test and all subsequent tests they took" and "students may report one or another nickname on follow-up tests." However, this is inconsistent with an account given to us by MPR, which implemented the experiment. Here is MPR's description:

---

[3] In an email correspondence dated Sept. 13, 2002, Howell explained the origin of their mistake: "I was working off of the uncorrected baseline weights."

> When a family came to the table we asked the mother's name, put the mother's label on her questionnaire and the kids labels on their questionnaires and test sheets. If a kid was in kindergarten his/her label just stayed on the list. ….in all four years the procedure was basically the same for all grades: mother to family to kids.[4]

In the judgment of John Homrighausen, the project's survey director, "both those with missing baseline scores and those that have baseline scores had an equal chance to be matched to their subsequent scores." Thus, we see no reason to put any stock in PH's second argument.

PH argue that nonrandom attrition could corrupt the experimental design. It could. Our paper and HP (2003), however, suggest that nonrandom attrition was not a problem for those with baseline scores. The same procedures were used to collect data from those with missing baseline data, so there is no reason to suspect that attrition caused bias for those with missing baseline scores. The fact that treatments and controls – including those not given baseline tests -- were well balanced in terms of baseline variables (see our Table 1) suggests attrition did not corrupt the experimental design.[5] Furthermore, appropriately re-weighting the data or imputing missing data is the correct statistical solution to nonrandom attrition (see Hausman and Spence, 1977), not simply controlling for the scores in a regression estimated for those with available data.

Notably, when Howell and Peterson evaluated parental responses, such as school satisfaction, they did *not* exclude students with missing baseline scores. Nor did they caution that their findings were less reliable as a result. If they did not trust the data or randomization for students with missing baseline scores, why did they include them in the analysis of parental outcomes?

We would ask readers to consider the following thought experiment: Is the voucher research influential because baseline scores were collected for a subset of students or because random assignment was used to assign students to treatment and control groups? We think it is the latter.

---

[4] Email correspondence, John Homrighausen, June 12, 2003.

[5] Students with and without baseline scores were matched to each other in the matched propensity design subsample, so we do not see how throwing out students with missing baseline scores improves the balance of the remaining sample, as PH argue. This argument certainly does not apply to the randomized block half of the sample.

If so, then the fact that the exclusion of students without baseline scores qualitatively changes the results is problematic.

When baseline scores are excluded and the full sample of Black students (using PH's racial/ethnic classifications) with follow-up scores is analyzed, the effect of vouchers on composite scores is less than a tenth of a standard deviation and statistically insignificant at the 10% level in Years 2 and 3 (see our Table 3b, column 3).[6] No further changes are needed to suggest that the original HP results are non-robust.

Race

PH are most acidic when it comes to race. They accuse us of using a flawed definition of race – actually, they mysteriously say "flawed ethnic classification scheme," even though their comments concern Black/African Americans and their questionnaire used the term "racial/ethnic categories." Apparently, they are trying to claim that they meant all along to treat African Americans as an ethnic group, distinct from Hispanics -- and they imply that we did too. But many Hispanics also consider themselves to be African American, and vice versa. According to Richard Rodriguez, "In Latin America, one sees every race of the world. One sees white Hispanics, one sees black Hispanics… To compare blacks and Hispanics, therefore, is to construct a fallacious equation." We explicitly allowed people to be in *multiple* racial/ethnic categories (see our Tables 5 and 6) to explore the robustness of their results.

PH disclose for the first time that they restricted their sample of African Americans to children of non-Hispanic, African American mothers. Their past accounts were vague or misleading. For example, Howell, et al. (2000) write, "If the trend line observed over the first two years continues in subsequent years, the black-white test gap could be eliminated in subsequent years of education for black students who use a voucher to switch from public to private school." It is unclear from this spectacular claim that the authors meant non-Hispanic,

Blacks. In their book, HP never state that they focus on non-Hispanic, Black/African Americans, as opposed to all Black/African Americans.

Of course, examining a sample of non-Hispanic, Blacks makes sense for some purposes -- but so does a sample of all Black/African Americans, regardless of nationality.[7] Our goal was to use the data to see if the results could generalize to the *entire* population of Black children, a sample that the reader might have inferred HP were interested in too. The question we asked is whether their already-fragile findings hold up if the sample is expanded to include children with a Black, Hispanic parent (who, incidentally, might also have another parent who is Black, non-Hispanic).

It is now manifestly clear that the racial/ethnic assignment of students in their original work was not well thought out, and the coding of race/ethnicity did not adhere to any consistent set of principles. Here's an example. PH write, "if respondents checked the 'other' category but then claimed to be 'Hispanic,' it seems safe to assume that they overlooked the Hispanic category above, making reclassification appropriate." But PH often violated this rule. Consider the write-in responses of each of the following cases: no. 14681 "Ecudor Hispana"; no. 16521 "Hispana"; no. 13543 "Hispanic Domini"; no. 7718 "Colombia"; no.'s and 27 and 219 "Honduras"; no.'s 3154 and. 585 "Mexican"; no. 14378 "Ecuador" -- all of these students were classified as *non-Hispanic*, and many more examples can be found.

Obviously, many write-in responses were overlooked by the coders. This lack of care is understandable once one recognizes that race/ethnicity were never intended to be a focus of the study. It was only happened upon after results for other subsamples were disappointing.

PH contend that no "clear" mistakes were made classifying African American students based on write-in responses. Again, we disagree. For example, case id 19651 was classified as

---

[6] This is not a matter of the baseline scores changing the results, as columns 1 and 2 of our Table 3b show that omitting baseline scores is inconsequential.

[7] Contrary to PH's claim, many African Americans who immigrated to the United States from Latin America also had a heritage of slavery and segregation.

non-Black but wrote-in "Africa" *and* checked the box indicating Black/African American (nonHispanic) in the follow-up years, and had test scores available. Other students were classified as non-Black when they had a Black/African American (non-Hispanic) father and *no data* reported on their mother.

Despite these problems, PH defend their racial/ethnic classification based on Q25 as "certainly consistent" with the OMB guidelines. We asked the office at OMB responsible for the standard for a judgment. OMB's response follows:

> Q25 is not in compliance with the 1997 Standards for Data on Race and Ethnicity (no longer referred to as Directive 15) for several reasons. First, it has a restriction to mark only one race and the standards are clear that multiple race reporting should be permitted. Second, the categories used are not in compliance .... Third, while the standard does permit a 'combined' race and ethnicity question such as Q25 on your instrument, two separate questions are preferred -- one that asks ethnicity and a second that asks race.[8]

When we asked whether it was appropriate under OMB guidelines for PH's combined question to instruct "MARK ONLY ONE" category, we were told: "whether it is combined or not, individuals should be able to select one or more races or races/ethnicity."

PH defend forcing individuals to make a choice between their race and Hispanic ethnicity when the OMB guideline is contrary. Moreover, some respondents are specifically directed away from choosing Black/African American, even if that was the category they (or their children) most closely identified with, because of the "non-Hispanic" parenthetical. Our point, which PH distort, was not that Sammy Sosa, a Dominican, should be classified as Black because of his skin color, but that Mr. Sosa would not have been given the opportunity to classify himself as Black, or as Black and Dominican, if he so chose, by PH's restrictive question and classification scheme.

A reasonable interpretation is that people who were unsatisfied with the constrained choices they were presented and wrote in "Black and Hispanic" intended to be classified as *both* Black and Hispanic, not as neither.

---

[8] Email correspondence, Susan Schechter, Senior Statistician, OMB, June 19, 2003.

PH's claim that "it is misleading—and contrary to the very federal guidelines that Krueger and Zhu use to bolster their case—to classify as 'Black, non-Hispanic' people who openly identify themselves as 'Hispanic,' 'Dominican,' or 'West Indian'" is unfair and incorrect. We did not classify anyone who wrote "Hispanic," "Dominican" or "West Indian" as "Black, non-Hispanic." Our Table 5 clearly says "for all Black students." If someone marked other and wrote in Hispanic/Black, we appropriately classified that person in both categories, *consistent with the OMB guidelines* and consistent with the individual's response. In their analysis, by contrast, PH failed to classify people who wrote "Dominican" or "Hispana" as Hispanic.

PH dismiss the role of Black fathers in unmarried families, despite evidence that many unwed Black fathers play an active role in childrearing (see Carlson and McLanahan, 2002). They also disregard the fact that, for better or worse, for centuries in the United States children were defined as Black if *one* of their ancestors was Black (see Davis, 2001, and Myrdal, 1944). PH's scheme would classify Halle Berry as non-Black -- a label she has rejected -- because she was born to a White mother and Black father. Our preferred approach would have been to allow students to self-identify their own race *and* ethnicity, or to have asked the parents to report the race *and* ethnicity their child identified with. Unfortunately, PH failed to collect such data.

Regardless, PH's (2003b) criticism that "[o]nly when Hispanics are defined as African-American or the data are loaded with meaningless extra information do effects attenuate" is factually incorrect. *For Year 3, for example, if we exclude all observations with write-in responses of "Black/Hispanic" and do not control for any covariates other than randomization strata, the voucher effect is only 1.63 points, with an insignificant t-ratio of 1.01.* This estimate is actually *smaller* than the comparable estimate in column 2 of our Table 5, so PH's comments about mixed heritage driving our results are misleading.[9]

---

[9] Because we and MPR could not replicate PH's samples in Table 3 and 4, and they did not cooperate in an attempt to understand the differences, we recommend skepticism concerning their analysis.

Specification

PH accuse us of "rummaging barefoot through the data." In fact, we presented a range of estimates, many with no covariates at all other than those that defined the random assignment, so the reader could judge the results for him or herself. We also used a rule that selected covariates to minimize the standard errors of the estimates. PH are displeased that the most precise estimates are insignificant, but that is not a consequence of our procedure for selecting covariates. This rule is just a more efficient version of PH's assertion that baseline scores should be controlled because they are highly correlated with follow-up scores. PH do not claim to have found an unbiased estimate that is more precisely estimated than the ones we present in Table 4.

Anyway, we are happy to throw out the covariates – the results are still insignificant (see the *simple and transparent* models in our Table 3b, column 3 and Table 5, column 2). PH resort to using a 20% threshold for a two-tailed significance test to argue that the results are still significant without covariates, but we do not expect many readers to be persuaded by this yardstick.[10]

PH selectively cite their research proposal, which they claim establishes a pre-specified plan to include baseline test scores. Their November 1997 proposal, however, stated: "We will estimate program impacts by comparing outcomes for students and parents in the scholarship program, with outcomes for those in the control group." A footnote attached to this sentence clarifies: "Additional analyses will compare achievement test score trajectories for the treatment and control groups. This will show how change in students' achievement is affected by the scholarship offers." One can infer that the primary plan was to present estimates of level effects, without controlling for baseline scores; analyses of changes in scores were to be ancillary.

Regardless, the beauty of a proper randomized experiment is that control variables are unnecessary. There is no evidence suggesting that attrition or other threats corrupted the integrity

---

[10] PH report this result as significant at the 10% level for a *one-tailed* test, but because they found a negative effect of vouchers in Washington, D.C. after 3 years, we see no reason to conduct a one-tailed test.

of the experimental design. Theory is not needed to determine which covariates to include, or what functional form they should take.[11] The estimates should be the same regardless of which baseline covariates are included. The fact that they move around so much – falling to insignificance when covariates are omitted, attaining statistical significance when the sample is limited to those with baseline scores, and falling to insignificance when a broader sample of Black students is analyzed – strongly suggests that the original HP results were spurious.

One should also recognize that an aspect of HP's specification search involved the sample. The aforementioned research proposal described several subsamples that would be analyzed; it never mentioned race. (Three primary subgroups were identified: multi-children families; students from below-average schools; students from religious families. Additionally, separate impact estimates were promised by grade level, baseline scores and parents' education.) Results in the first year report were *not* broken down by race (Peterson, et al. 1998). Only after the other samples showed no impact were results tabulated by race. If enough subgroups are examined, one is bound to exhibit a significant effect, although it is unlikely to be robust.

Moreover, their proposal makes no mention of PH's controversial claim that the literature finds voucher effects for Black students but not others. PH's reference to their other experiments yields no support for this claim either, as they find that vouchers had a small *negative* effect on Black students' scores after three years in the D.C. experiment, yet they overlooked this finding and only reported the second year effect. And Peterson and Greene did not report separate estimates by race for Milwaukee. Had race and ethnicity been intended to be a focus of the study – rather than something arrived at through specification searching – surely they would have collected data on the race and ethnicity of the students.

---

[11] PH's claim that a high correlation between baseline scores and subsequent scores is theoretical justification for their inclusion entails a novel conception of theory. Usually theory involves assumptions and predictions, not a statement of a correlation. See Todd and Wolpin (2003) for a critical theoretical appraisal of models that control for initial test scores.

Robustness

PH (2003a) claim that "108 of 144 different statistical models" that they estimated "yield positive and significant effects," and that this shows their results to be robust. This claim has several problems. First, the data are not independent, so simply counting positive, significant results is a misleading indicator of the strength of the results. Second, the order in which one adds additional covariates to a series of multiple regressions affects the significance of the estimates at each step. If we started with a different set of explanatory variables and added more controls, we could generate more of fewer insignificant results. Third, PH (2003b) claim that 108 of 120 estimates were significant, so obviously the results are dependent on the models they chose to estimate -- and 0 of 24 of their new models yielded significant results. Fourth, most of their estimates use their original problematic sample that unnecessarily throws out observations. Fifth, we could not replicate their results: when we estimate their models we find 74 of 144 estimates were significant at the 5% level; for Year 3, only 23 of 48 were significant.

PH's exercise actually shows a puzzling lack of consistency in the results, not evidence of robustness. HP (p. 53) present a power calculation for Year 3 that implies that more than 80 percent of (independent) estimates for African Americans would be significant at the 5% level if the true effect was as large as they claim (even after knocking it down by 10 percent because of the weights and strata controls). Even with their dependent estimates – which would raise the expected figure above 80% -- they are well below that threshold.

PH further assert that our results show theirs to be robust, claiming that "30 of 51 of [KZ's] estimations of the voucher impacts for Blacks are significantly positive." This claim has two problems. First, many of our estimates were just an attempt to replicate the sample, weights and strata controls in their earlier work, which they now acknowledge were flawed. Only 7 of the 27 of our estimates that used a different sample than their narrow one were statistically significant at

the 5% level. Second, the estimates are not independent, so again this vote counting exercise is misleading. This hardly qualifies as a robust effect.

Two-Stage Least Squares

PH continue to report 2SLS estimates for attending private school for three full years, despite the likely upward biases from partial compliance (i.e., some voucher students only attended private school for one or two years) and measurement errors in their public/private school enrollment data. They do not respond to our criticisms of their 2SLS model, and they do not comment on our alternative 2SLS model (which uses years in private school as the endogenous variable).

To defend their 2SLS models, HP reference Krueger's (1999) study of the Tennessee STAR class-size experiment for a "similar use of 2SLS." The similarities are not very great, however. Krueger does not instrument for attending a small class for all three years (though he could have), but just for actual class size in any particular year. And unlike PH, Krueger reports OLS estimates of treatment effects as well as 2SLS models. Anyway, the OLS estimates we reported highlight what is at issue in this exchange – whether chance can be ruled out as the explanation for the small difference in scores between those offered and those not offered a voucher.

Conclusion

PH's contention that two of three changes to their analysis must be "privileged" to cast doubt on their original results for African American students is misleading. If one starts with their sample and simply adds in students with missing baseline scores – and does not control for any covariates other than randomization strata – the t-ratio for the effect of offering vouchers on the Year 3 composite score falls from 3.416 to 1.635. In year 2, which PH emphasize for D.C. and Dayton, the corresponding t-ratio is just 1.46. A 95 percent confidence interval for the voucher effect runs from -0.8 to 5.8 points in Year 2 and from -0.6 to 6.1 points in Year 3. The conclusion that vouchers had no effect on achievement cannot be rejected at a 10% (or lower) level of

significance, even using their incomplete sample of children of Black/African American (non-Hispanic) mothers.

If one adds controls for a variety of covariates *or* expands the sample of African American students to include those with an African American father – legitimate checks of robustness – the results point to even smaller effects that easily could have occurred by chance. The most precise estimates are small and statistically insignificant. These are hardly robust findings, especially in light of the fact that the standard errors do not account for additional variability due to the estimation of follow-up weights or the search PH conducted over the subsample.

The most plausible reconciliation of the results is that when they started PH expected vouchers to raise scores for all inner-city students.[12] When this did not materialize in any of their three cities, PH sought a subsample for which they could claim vouchers raised achievement. This led them to a relatively small and oddly defined subsample of non-Hispanic, Black students with initial test scores. But the result was spurious, so it did not hold up: after three years in D.C. it fell apart; it was never very strong in Datyon; and we have shown that reasonable changes to the sample render the voucher effect statistically insignificant and small in New York City.

In their thoughtful response to our paper, Mayer and Myers (2003), PH's original collaborators, wrote, "the new evidence presented by Krueger and Zhu suggests that one must remain cautious when interpreting the findings for African Americans." PH's response does nothing to dispel the required cautions. Because PH did not collect or classify students' racial/ethnic data in an accepted manner, and because their results for African American students do not hold up under plausible changes to the sample or statistical model, doubts will always linger about their findings. Nevertheless, PH's experiments convincingly show that poor minority students as a group did not experience achievement gains from vouchers.

---

[12] For example, shortly before the experiments were launched, Peterson (1995) proclaimed: "When religious issues are ignored and vouchers are evaluated in educational terms, the findings are clear: Vouchers work for the inner city poor. … Those [voucher students] attending the (mostly Catholic) schools participating in New York's SSP program made remarkable test score gains."

References

Carlson, Marcia and Sara McLanahan, "Early Father Involvement in Fragile Families", Center for Research on Child Wellbeing Working Paper, Princeton University, November 2002

Davis, James, *Who Is Black?* Penn State University Press, 2001.

Jay Greene and Paul Peterson, "Methodological Issues in Evaluation Research: The Milwaukee School Choice Plan," mimeo, August 1996. www.ksg.harvard.edu/pepg.

Hausman, Jerry, and A. Michael Spence, "Non-Random Missing Data," MIT Working Paper 200, May 1977.

Howell, William and Paul Peterson, with Patrick Wolf and David Campbell, *The Education Gap: Vouchers and Urban Schools*, Washington, DC: Brookings Institute Press, 2002 (Advance Reading Copy).

Howell, William and Paul Peterson, "Uses of Theory in Randomized Field Trials: Lessons from School Voucher Research on Disaggregation, Missing Data, and the Generalization of Findings," mimeo, August 2003.

Howell, William, Patrick Wolf, Paul Peterson and David Campbell, "Test-Score Effects of School Vouchers in Dayton, Ohio, New York City, and Washington, D.C.: Evidence from Randomized Field Trials," Program on Education Policy and Governance, Harvard University, http://www.ksg.harvard.edu/pepg/pdf/dnw00x.pdf, August 2000.

Myers, David and Daniel P. Mayer, "Comments on 'Another Look at the New York City Voucher Experiment'," mimeo., April 1, 2003.

Myrdal, Gunnar, *An American Dilemma: The Negro Problem and Modern Democracy*, New York: McGraw-Hill,1944.

Peters, Paul, "Vouching for a Religious Education," Op-Ed, *The Wall Street Journal*, December 28, 1995.

Peterson, Paul and William Howell, "Efficiency, Bias, and Classification Schemes: Estimating Private-School Impacts on Test Scores in the New York City Voucher Experiment," mimeo, 2003a.

Peterson, Paul and William Howell, "School Vouchers Work," Letter to the Editor, *New York Times*, June 20, 2003b.

Peterson, Paul and William Howell, "A Further Look at the New York School Voucher Randomized Field Trial: Comments on Alan B.Krueger and Pei Zhu's 'Another Look at the New York School Voucher Experiment'," Paper presented at the Conference on Randomized Experimentation in the Social Sciences, Yale University, August 2002.

Peterson, Paul, David Myers, and William G. Howell, "An evaluation of the New York City School Choice Scholarships Program: The First Year," Mathematica Policy Research and Program on Education Policy and Governance, Harvard University, October, 1998.

Rodriguez, Richard, "'Blaxicans' and Other Reinvented Americans," *The Chronicle of Higher Education*, September 12, 2003.

Todd, Petra and Kenneth Wolpin, "On The Specification and Estimation of The Production Function for Cognitive Achievement," *Economic Journal*, 2003, vol. 113, issue 485, pages F3-F33.