

A Reply to “Critique of ‘An Evaluation of the Florida A-Plus Accountability and School Choice Program’” by Gregory Camilli and Katrina Bulkley in Education Policy Analysis Archives, Volume 9 Number 7 March 4, 2001, <http://epaa.asu.edu/epaa/v9n7/>

By Jay P. Greene

**Senior Fellow, The Manhattan Institute for Policy Research
March 5, 2001**

The Camilli and Bulkley re-analysis of my evaluation of the Florida A-Plus choice and accountability program is neither factually nor conceptually accurate. They mischaracterize my findings to create a straw man that is easier to knock down. They then compare scores for different samples across time to bias the results downward. They follow this by disaggregating results by grade level to produce smaller samples, making the detection of significant effects more difficult. They then propose a correction for regression to the mean that absorbs much of the effect that could reasonably be attributed to the prospect of vouchers, making the effect sizes even smaller. And finally they measure effects for school level results as a proportion of estimated student level standard deviations to ensure that the positive results they have found (despite their best efforts) are made to seem ridiculously small. The Camilli and Bulkley re-analysis is almost a textbook for how to do a hatchet job on positive results that one wishes to make go away.

First, they build the straw man. They attribute to me the claim that the effect of vouchers in Florida was between .80 and 2.23 standard deviations. They say: “These gains for ‘F’ schools were then translated into effect sizes for reading (.80), mathematics (1.25), and writing (2.23) (Greene, 2001a, endnotes 12-14). No doubt, as computed, these gains are statistically significant. They are also among the highest gains ever recorded for an educational intervention. Results like these, if true, would be nothing short of miraculous, far outpacing the reported achievement gains in Texas and North Carolina.”

I obviously did not claim these as the effect sizes for vouchers. These numbers were taken from endnotes that simply described how large the year-to-year changes for F schools were. My claims for voucher effects are clearly described in the text as: “The improvement on the reading FCAT attributable to the prospect of vouchers was a modest 0.12 standard deviations and fell short of being statistically significant. The voucher effect on math scores was a larger 0.30 standard deviations, which was statistically significant. And the prospect of vouchers improved school performance on the writing test by 0.41 standard deviations, an effect that is also statistically significant.” (p. 8) These effects were also clearly displayed in Table 3 and marked as “Voucher Effect Measured In Standard Deviations.”

Second, they appear to have compared scores for different samples over time in a way that biases certain results downward. They are no more accurate in describing the test score data available on the Florida Department of Education (FDOE) web site, which

they used for their own analyses, than they were in describing my findings. They say: “An alternative method of choosing a sample is to use the results for all curriculum groups, and these data are available on the Florida Department of Education web pages.” This is not correct. The FDOE web site (<http://www.firn.edu/doe/sas/fcdtrpt00.htm>) contains scores for all curriculum students for 2000 but only standard curriculum students for 1999. Comparing scores for these two different samples, as Camilli and Bulkley appear to have done, biases all gains downward because exceptional education students who were excluded in 1999 were included in 2000. My analysis of scores from standard curriculum students in both years is an apple-to-apple comparison.

While their apple-orange comparison biases all gains downward, it does not fundamentally distort the gains of F schools relative to other schools. As I observed in endnote 10 of my report: “the web site only has scores for standard curriculum students in 1999 and all students in 2000. This study used scores for standard curriculum students in both years. Earlier analyses on these results from the web site do not produce results that are substantively different from those reported here. This suggests that the inclusion or exclusion of test scores from special needs students has little bearing on the conclusions of this evaluation.” Given how much attention Camilli and Bulkley appear to pay to endnotes, one would have thought that this would have addressed their concern about whether the sample should have included all curriculum students or only standard curriculum students.

Third, they make the case for disaggregating the results by grade level, the net effect of which is to produce smaller samples and less stable results. Their argument for disaggregating is based largely on the results in their Table 2, which purport to show that the year-to-year average changes in FCAT scores differ for different grade levels. Remember that their test score analyses are using an apple-orange comparison of standard curriculum students in 1999 to all curriculum students in 2000, so the changes in test scores reported in their Table 2 are incorrect and all biased downward. Also note that disaggregating by grade level produces samples with only a handful of F schools in grades 8 and 10, making any findings about the progress of schools in those grades that faced the prospect of vouchers unstable and insignificant.

In addition to relying on incorrect comparisons of 1999 and 2000 test scores by grade level, Camilli and Bulkley make a theoretical argument for disaggregating results. They argue: “the results of a policy implementation may be different at different grades, even if this is not an a priori expectation.” The results of policy implementation may be different in rural and urban areas. Why not disaggregate the results by grade level and urbanicity? The results of policy implementation may also be different in each of the 64 school districts in Florida. We could disaggregate to incredibly small samples if we wished. The obvious argument against disaggregating results, even when plausible differences in policy implementation may exist, is that we do not want to disaggregate results so that samples are too small to be reliable unless there is compelling evidence that requires disaggregation. Disaggregating by grade level gives us only a handful of failing schools to examine in grades 8 and 10; numbers that are too small to yield reliable

results. Since Camilli and Bulkley do not provide compelling evidence for disaggregating results into such small units, we ought not to do so.

Fourth, Camilli and Bulkley propose an alternative, and biased, way of addressing the possibility that regression to the mean might account for at least some of the improvement at schools that faced the prospect of vouchers. Their proposal is the effect of regression to the mean could be modeled as the slope of the regression line produced by estimating this year's scores based on last year's scores. The "true" gain for F schools would then only be the amount by which F schools improved beyond the improvement predicted by the regression line. That is, in their view the voucher effect can accurately be measured as the error term from the regression model.

Among the many problems with the approach they propose for correcting for regression to the mean is that their estimate of the influence of regression to the mean, the slope of the regression line, is actually influenced by the magnitude of the true voucher effect. Let's say that we found a program with the same amount of regression to the mean as in Florida but the true voucher effect were twice as large. If we then introduced the "correction" that Camilli and Bulkley propose we would wrongly attribute some of the true voucher effect to their adjustment for regression to the mean and underestimate the voucher effect. This would occur because the slope of the line that predicts one year's scores from the previous scores would become more steep because the failing schools experienced larger true gains. Remember under our hypothetical the influence of regression to the mean is unchanged yet their "correction" for regression to the mean would change with a change in the true voucher effect.

Obviously a correction for an error that increases when the true effect increases is seriously flawed because it is counting as error effects that are true effects. Camilli and Bulkley's correction for regression to the mean is similarly seriously flawed because the line that failing schools must over-perform in their analysis is itself influenced by the size of the true voucher effect among failing schools.

A more reasonable correction for regression to the mean would estimate the slope of the line for predicting one year's scores from the previous year's scores excluding the failing schools and then judge the extent to which the schools that faced the prospects of vouchers over-performed that expectation. This way the correction for regression to the mean could not change under a scenario in which the true voucher effect was made larger while the true regression to the mean remained constant. I performed precisely this type of analysis in Table 5 to show that the true voucher effect is around 4 points in reading, 8 points in math, and .25 points on writing (which uses a different scale).

Fifth, Camilli and Bulkley wish to covert all of these point gains into smaller changes in terms of standard deviations by using an estimated student-level standard deviation instead of the school-level standard deviation that I use. They do not describe how they estimate the standard deviation for individual student test scores, but it is 3.5 times as large as the standard deviation for the results reported as school averages. It is true that variation in the scores will be greater at the student level than at the school level,

but it is not at all clear why one should use individual level variation for calculating effect sizes.

The unit of analysis in my study is correctly the school level. Schools are assigned grades by the state, not students. Schools face the prospect that their students will be offered vouchers if they do not improve. Schools must develop strategies for improving their efforts so that test scores will rise. The results in my study are appropriately reported as school averages and the effect sizes are appropriately computed using school level variation in scores. The only obvious appeal for using student level variation to compute effect sizes is that it makes those effects three and one half times smaller.

By this time it should be clear that making the positive effects from the A-Plus choice and accountability program smaller was probably the point of Camilli and Bulkley writing their piece, not attempting to identify the program's true effects. At each point they distorted my findings, misrepresented the data, or employed analytical techniques that appear designed to minimize the positive results from the A-Plus program.

Let me quickly note the other important finding of my report that they did not challenge: the Florida Comprehensive Assessment Test (FCAT) is a reliable measure of student performance because its results correlate highly with the results of a low-stakes standardized test administered around the same time. They note that the correlation at the school level is higher than it would be at the individual level (as I did in endnote 11), but they do not challenge the claim that Florida's testing program produces reliable measures of student performance. This implicit concession on their part is important because the finding contradicts the claims of opponents of testing and accountability programs who regularly appear in the web-only Education Policy Analysis Archives.