Education Policy Analysis Archives

Volume 9 Number 7 **March 4, 2001** **ISSN 1068-2341**

A peer-reviewed scholarly journal
Editor: Gene V Glass, College of Education
Arizona State University

Articles appearing in **EPAA** are abstracted in the *Current Index to Journals in Education* by the ERIC Clearinghouse on Assessment and Evaluation and are permanently archived in *Resources in Education*.

Critique of "An Evaluation of the Florida A-Plus Accountability and School Choice Program"

Gregory Camilli
Rutgers University

Katrina Bulkley
Rutgers University

Related article.

Abstract
In 1999, Florida adopted the "A-Plus" accountability system, which included a provision that allowed students in certain low-performing schools to receive school vouchers. In a recently released report, *An Evaluation of the Florida A-Plus Accountability and School Choice Program* (Greene, 2001a), the author argued that early evidence from this program strongly implies that the program has led to significant improvement on test scores in schools threatened with vouchers. However, a careful analysis of Greene's findings and the Florida data suggests that these strong effects may be largely due to sample selection,

> regression to the mean, and problems related to the aggregation of test score results.

One of the most closely watched state reforms in recent years is the use of school vouchers as a part of the accountability system for Florida's public schools. This program is of particular interest because of its strong similarities with proposals put forward by President George W. Bush. As a New York Times article noted, "Gov. Jeb Bush's educational program in Florida has been held up as a model for its combination of aggressive testing of schools' performance, backed by taxpayer-financed vouchers, which his brother President Bush is proposing for the nation as a whole" (Schemo, 2001).

A recently published report purports to show a convincing link between the threat of school vouchers for students in certain low-performing schools in Florida and achievement gains in those schools. *An Evaluation of the Florida A-Plus Accountability and School Choice Program* (Greene, 2001a) documents gains in achievement on the Florida Comprehensive Assessment Test (FCAT) in the areas of reading, mathematics, and writing. (This evaluation will be referred to as *Evaluation of Florida's A-Plus Program*, for short.) These findings, not surprisingly, have received a substantial amount of attention in the popular press (cf. Schemo, 2001; Lopez, 2001; Greene, 2001b). The gains reported are attributed to incentives implemented under Title XVI (section 229.0535 "Authority to enforce school improvement") of the 2000 Florida Statutes:

> It is the intent of the Legislature that all public schools be held accountable for students performing at acceptable levels. A system of school improvement and accountability that assesses student performance by school, identifies schools in which students are not making adequate progress toward state standards, institutes appropriate measures for enforcing improvement, and provides rewards and sanctions based on performance shall be the responsibility of the State Board of Education.

In the A- Plus accountability system, schools are evaluated and assigned one of five grades (A, B, C, D, F) based primarily on FCAT scores, and to a lesser extent, the percent of eligible students tested and dropout rates (Florida Department of Education, 2001). If a school receives two grades of "F" in any four-year period, it becomes eligible for state board action. Contrary to the implication in Greene's title, such action is not limited to school choice; rather, actions may include providing additional resources, implementing a school plan or reorganization, hiring a new principal or staff, and other unspecified remedies designed to improve performance. However, the possibility of public schools losing children to either private schools or higher-performing public schools is clearly the area of most interest and controversy. In the 1999-2000 school year, two Pensacola elementary schools met the eligibility criteria (Note 1), and as a result, lost 53 children to private schools and 85 to other public schools.

Greene argued that his report "shows that the performance of students on academic tests improves when public schools are faced with the prospect that their students will receive vouchers" (p. 2). At the center of his argument is the fact that all 78 schools that received an "F" in 1999 received a higher grade in 2000. His claim that the threat of vouchers was responsible for the improvement of "F" schools (from the 1998-1999 to the 1999-2000 school year) includes several important elements. First, an attempt was made to show the validity of the FCAT by showing a strong correlation to another test (Stanford-9) given in Florida in 2000. Given this evidence, he then

proceeded to show the average gains for each school receiving a particular grade. Based on the latter results, it was concluded that:

> The most obvious explanation for these findings is that an accountability system with vouchers as the sanction for repeated failure really motivates schools to improve. (p. 9)

However, Greene also wrote:

> While the evidence presented in the report supports the claims of advocates of an accountability system and advocates of choice and competition in education, the results cannot be considered definitive. (p. 9)

The A-Plus accountability system was duly noted as being relatively new, with the voucher options used in only two schools in the state, and possible—though not likely—manipulation of FCAT scores. It is an additional alternative that Greene mentions, commonly known as *regression to the mean*, that is one main concern of this report. This paper also examines three other issues: (1) sample selection, (2) the combining of gain scores across grade levels, and (3) the use of the school as the unit of analysis. Below, we subsume the latter two items under the category of "aggregation."

The potential policy importance of the findings Greene reports places a heavy burden on his study to demonstrate that the improved scores in schools that had previously received one "F" are in fact meaningful improvements and a result of school changes linked to the threat of vouchers. We argue here that the evidence does not support this conclusion. We show that there may have been some small achievement gains in Florida from 1999-2000, but these effects were vastly overestimated in Greene's analysis. However, even if these modest outcomes withstand further investigation, it is not at all clear that they resulted from the threat of vouchers as opposed to other aspects of the accountability program.

Background

Several recent reforms have similar components to the Florida effort. It is not the purpose of this report to review that literature, but two well-known reforms deserve mention. One of these, which Greene specifically addresses, is the Texas accountability system and its use of the Texas Assessment of Academic Skills (TAAS). Another is the public voucher program in the city of Milwaukee. Comparisons between each of these reforms and the Florida's A-Plus accountability system are limited for a variety of reasons. The accountability system in Texas varies in critical ways from the model in Florida, especially in the use of vouchers as a sanction in the latter state but not the former. Greene did, however, address an important methodological concern (discussed below) that arose in a recent study of the TAAS (Klein, Hamilton, McCaffrey, and Stecher, 2000). In the area of publicly-funded vouchers, students in Milwaukee who met certain income requirements are eligible to receive vouchers allowing them to attend local private schools. Several evaluations have been done of this program (i.e. Witte, 1996; Greene, Peterson and Du, 1998). These evaluations are not comparable to the Florida evaluation because they examined the test scores of individual students who either received vouchers or applied for vouchers but did not receive one; the Greene study focuses on the school impact on test scores of the threat of vouchers, not the actual

provision of vouchers.

Summary of the *Evaluation of Florida's A-Plus Program*

In *Evaluation of Florida's A-Plus Program* (Greene, 2001a, Table 2), the main results were obtained by aggregating across grade for school types A, B, C, D, and F. These results are reproduced in Table 1 below.

Table 1
FCAT Reading and Mathematics 1999-2000 Gains from Greene's "An Evaluation of the Florida A-Plus Accountability and School Choice Program"

Grade	Reading	Math	Writing
A	1.90	11.02	.36
B	4.85	9.30	.39
C	4.60	11.81	.45
D	10.62	16.06	.52
F	17.59	25.66	.87

To obtain the overall reading and writing gain, gains at the 4th, 8th, and 10th grade levels were pooled, while for mathematics, gains at the 5th, 8th, and 10th grade levels were pooled. School means for standard curriculum students were used to compute gains, not individual student scores. It can be seen that the average gain for "F" schools "are more than twice as large as those experienced in schools with higher state-assigned grades" (Greene, 2001a, p. 6). 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. This may have moved Greene to conclude:

> While one cannot anticipate or rule out all plausible alternative explanations for the findings reported in this study, one should follow the general advice to expect horses when one hears hoof beats, not zebras. The most plausible interpretation of the evidence is that the Florida A-Plus system relies upon a valid system of testing and produces the desired incentives to failing schools to improve their performance. (p. 14)

Critique of the *Evaluation of Florida's A-Plus Program*

Our critique of Greene's evaluation focuses primarily on two problematic issues: aggregation and regression to the mean. We do not examine in detail Greene's validation argument for the FCAT based on its correlations with the Stanford-9 (the latter given in

2000). Greene's correlational analysis was conducted partly in response to concerns raised by Klein and his colleagues (2000) about the validity of the TAAS in Texas. However, it is worth noting that while the two tests have substantial correlations (in the range .85-.95), correlation coefficients computed on aggregate scores typically have much higher values than those computed with student scores. For example, school means on the reading and mathematics sections of the FCAT in 8th grade have a correlation of about .96. This correlation should *not* be interpreted as meaning that the FCAT reading and mathematics tests are statistically indistinguishable, but rather that correlations on aggregate score tend to be much higher than those for individual scores.

Sample Selection

Greene (2001a) used the school means of "standard curriculum" students to obtain school-level gains scores. Here "standard" defines a subset of students who tend to score higher on the FCAT (i.e., it does not include certain types of students with disabilities). 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. While there is nothing intrinsically wrong with using standard curriculum students, for the purposes of evaluation, however, it would seem preferable to look at the potential impact of the A-Plus program on all curriculum groups. Florida administrative statues allow for (or require) nontrivial variation in populations selected for determining school grades (Note 2).

Aggregation

In the analyses below, we disaggregate results by grade. This is useful because overall state gains (Florida Department of Education, 2001) vary by grade as shown in Table 2.

Table 2
FCAT Score Gains from School Year
1998-1999 to 1999-2000

Grade	Reading	Math	Writing
4	5.0	N/A	0.0
5	N/A	11.0	N/A
8	-5.0	7.0	0.0
10	-4.0	3.0	0.1

The data in Table 2 suggest several problems with aggregation across grades. First, the results of a policy implementation may be different at different grades, even if this is not an a priori expectation. Second, in order to fine-tune a successful policy—or weed-out an unsuccessful policy—suitable diagnostic information is critical. Furthermore, a subtle problem arises when mixing the scales of two different instruments given at different grades. How can we be sure that this isn't the old apples-and-oranges problem? To be safe, the best advice is to conduct separate analyses and

then to combine them while making explicit the assumptions involved.

A more subtle problem involves the computation of effect size (Hedges, 1985), which is typically taken to be

$$\delta = \frac{x - E[x]}{\sigma}$$

This formula can be read as the difference between an observed value and an expectation divided by the standard deviation. In practice, the expectation ***E[x]*** could be a school's average test score for the prior year, and ***x*** could be taken as the score for the current year. It is also typical practice to use a measure of student individual variation in the denominator for "sigma" to facilitate a standard interpretation. For example, $\delta = 1$ means that the average student in the "treatment" population scores at the 84th percentile of the "control" population. Likewise, $\delta = 2$ means that the average student in the "treatment" population scores at the 98th percentile of the "control" population. So the interpretation is anchored in individual student achievement.

In contrast, Greene computed effect sizes relative to the standard deviation (SD) of schools, and though this is technically defensible, it must be recognized that such an effect size doesn't have the usual interpretation. In fact, we have estimated that the individual-level standard deviations (SD) are about 70 score points for reading and mathematics, and about .85 for writing—while the school-level SDs are about 20 points for reading and writing, and about .39 point for writing. Thus, an effect size for reading based on the school-level SD would be 350% larger than one based on the individual-level SD. At face value, the effect sizes computed by Greene, ranging from .80 to 2.23, are implausible because many studies have found that especially large educational effects (produced under laboratory conditions) fall into the range of .4 - .7.

But even if Greene's effect sizes are rescaled for comparability, they are still inflated by other factors including regression to the mean (see below) and an inappropriately selected definition of the expectation ***E[x]***. In regard to the latter issue, the effect of a treatment is usually defined as the net effect above and beyond average growth (the latter is referred to by statisticians as the *grand mean*). Thus, gain is defined as the net effect above average, and loss as the net effect below average. In this case, the average is the overall state gain; and the deviation from the grand mean represents the unique effect of a particular treatment or intervention. For example, take the average state gain for 4th grade reading in Table 2 of 5 points. If an intervention is defined as positive, it should register as being greater than 5 points since 5 points is what could be expected with no intervention whatsoever. It's not very useful to apply this correction to Greene's Table 1 because the results are aggregated across grades. However, in our analyses below, we build in this correction. We also use the individual-level standard deviation to facilitate the comparability of effect sizes to the general research literature.

Regression to the mean

Campbell & Stanley (1966) in their classic volume *Experimental and Quasi-Experimental Designs for Research* defined the *internal validity* of an experiment as:

> The basic minimum without which any experiment is uninterpretable: Did in fact the experimental treatments make a difference in this specific experimental instance? (p. 5)

In a very simple investigation, there are only two measurements taken: the pretest (O_1) and, after the experimental intervention, the posttest (O_2). Campbell and Stanley (1966) listed five definite weaknesses of this "One-Group Pretest-Posttest Design" and one potential concern which is of central importance to Greene's evaluation: *regression to the mean* or, alternatively, *regression artifacts*. They explained:

> If, for example, in a remediation experiment, students are picked for a special experimental treatment because they do particularly poorly on an achievement test (which becomes for them O_1), then on a subsequent testing using a parallel form or repeating the same test, O_2 for this group will almost surely average higher than O_1. This dependable result is not due to any genuine effect of [the intervention], and test-retest practice effect, etc. It is a rather tautological aspect of the imperfect correlation between O_1 and O_2. (p. 10)

In short, experimental units chosen on the basis of extreme scores tend to drift toward the mean upon posttest: low scores drift upward and high score drift downward. Campbell and Stanley (1966) then gave an extended treatment to this topic because "errors of inference due to overlooking regression effects have been so troublesome in educational research," and "the fundamental insight into their nature is so frequently missed" (p. 10). The regression phenomenon emerged from Francis Galton's studies of inheritance in biology, and this subject provides the most common phrasing of the regression to the mean effect: tall fathers tend to have tall sons, but not as tall on average as the fathers; while short fathers have short sons, but not as short on average as the fathers.

It can be seen in Table 1 for all three FCAT subjects that the trend is for higher achievement schools to gains less and lower achievement schools to gain more. This is a tell-tale sign of statistical regression, that is, scores in the tails of the distribution tend to drift toward the mean. Higher scores drift downward and lower scores drift upward relative to average gains. Greene (2001a) did consider this possibility, but rejected it as a potential explanation, arguing that:

> Regression to the mean is not a likely phenomenon for the exceptional improvement made by the F schools because the scores for those schools were nowhere near the bottom of the scale for possible results. The average F school reading score was 254.70 in 1999, far above the lowest possible score of 100.

Likewise, the average FCAT mathematics and writing scores of the F schools were 272.5 on a scale of 100-500 and 2.40 on a scale form 1-6, respectively. Greene thus concluded that regression to the mean was not a problem because the scores of the F schools were not at all extreme.

This is an inaccurate notion of regression to the mean because "extremeness" should be evaluated in terms of distance (in standard deviation units) below the overall group mean, rather than relative to the lowest possible score. A good measure of "distance below the mean" can be given in z-score units which are interpreted as "standard deviations below the mean" in the distribution of school means; z-scores of –3.00 and lower generally indicate substantial distance below the mean. To check for extremeness, we calculated the z-scores of the lowest performing school in 4th, 8th, and 10th grade reading, and 5th, 8th and 10th grade mathematics. These z-scores ranged

from a high of –3.2 to a low of –4.5, indicating a strong likelihood of obtaining a regression artifact in simple difference scores; however, the writing scores tended to be less extreme for the "F" schools.

In North Carolina, it was recognized that "Students who are proficient may grow faster" and "students who score low one year may score higher the next year, partly due to 'regression to the mean'" (Public Schools of North Carolina, 2000, p. 2). Both influences on achievement are explicitly taken into account in the North Carolina system when computing expected growth for schools. As noted by Campbell and Stanley (1966) the incorrect interpretation of regression effects has plagued educational research for decades. To give an example, consider a study by Glass and Robbins (1967) in which the SAT was given to a group of students, and researchers then took the high scorers as the control group and the low scorers as the treatment group. Predictably, the treatment showed a positive effect that disappeared when regression effects were taken into account (Glass & Robbins, 1967)

Methods

Data Sources

The state of Florida has an exceptional policy of granting the public full access to state, district, school level test scores, and other variable such as class size, per pupil expenditures, and the like. These data files containing school means for all curriculum students can be downloaded in the form of Excel spreadsheets at the Florida Department of Education website. For the present analysis, reading and mathematics, and writing FCAT scores at the school level were downloaded for both the 1998-1999 and 1999-2000 school years. Department staff provided a spreadsheet containing school grades, with district and school identification numbers, for the 1998-1999 school year.

Residual gain score analysis

Since we strongly suspected that the statistics in Table 1 were affected by at least two sources of error (regression to the mean and incorrect definition of net effect), we reanalyzed the data using the technique of *residual gain scores*. Glass and Hopkins (1996) described the context for residual gains:

> Administering parallel forms of the achievement test before [O_1] and after [O_2] instruction, then subtracting the pretest score from the posttest score [$O_2 - O_1$] for each student produces a measure that is far closer to the researcher's notion of a measurement of an achievement gain. One difficulty remains: Such a posttest-minus-pretest measure, [$O_2 - O_1$], is contaminated by the regression effect, usually correlate negatively with the pretest scores [O_1] … A better method to measure gain or change is to predict posttest scores [$\boldsymbol{O}_2$] from pretest scores [O_1] and use the deviation [$O_2 - \boldsymbol{O}_2$] as a measure of gain, above and beyond what is predictable by the pretest alone. (p. 167)

In the present case of the FCAT scores, O_1 is the pretest and O_2 is the posttest. Everything else in the present case is the same as in Glass and Hopkins's recommendation. By using residual gains, two goals are accomplished. First, the

regression effect is removed because the predicted score takes into account movement toward the mean. Second, the predicted value takes into account the average state gain; it will lead to unique net (policy) effects for any particular accountability grades.

Results

Average residual gains for the FCAT reading and mathematics tests, disaggregated by grade, are given in Tables 3 (reading), 4 (mathematics), and 5 (writing) below.

Table 3
Average Residual Gains for FCAT Reading

GRADE	GROUP	Mean	N	SD
4	A	1.45	121	8.30
	B	3.23	212	10.26
	C	-.86	694	10.54
	D	-.91	455	13.86
	F	2.35	66	12.96
8	A	.44	73	6.94
	B	1.03	90	7.68
	C	-.06	255	8.19
	D	-1.71	94	10.29
	F	7.26	7	12.84
10	A	.79	8	3.89
	B	2.55	12	4.77
	C	-.18	280	6.99
	D	.62	57	8.82
	F	-5.53	4	11.40

Table 4
Average Residual Gains for FCAT Mathematics

GRADE	GROUP	Mean	N	SD
5	A	4.30	121	7.61
	B	.17	210	8.71
	C	-.05	695	10.83
	D	-1.80	449	13.81
	F	4.36	66	15.13
8	A	.39	73	6.98
	B	.32	90	8.54
	C	-.19	255	9.28
	D	-.75	94	10.97
	F	8.78	7	10.82
10	A	1.88	8	2.47
	B	1.59	12	4.45
	C	-.18	280	6.66
	D	.78	57	8.91
	F	-6.73	4	14.73

In Tables 3 and 4, the largest effects are in the 8th grade, but in terms of standard deviation (SD) units, these effects are small (Note 3). Using the individual student SD of about 70 (versus the school SD of about 23), the effect size for 8th grade reading is δ = .10, and for 8th grade math is about δ = .13. We think it is not worthwhile to persevere on whether these effects are statistically significant because they are relatively small and other sources of possible bias cannot be plausibly ruled out as causes. For example, slight nonlinearities in the regressions might account for the higher effect sizes for the 8th grade F schools. In addition, the average effect for this group of only 7 schools is accompanied by a relatively high standard deviation. This means the overall positive effect is highly variable.

The results for FCAT writing are somewhat different for those in reading and mathematics. It can be seen in Table 5 at the 4th grade level that the average residual gain was .20 point on a scale that ranges from 1-6, and this effect is statistically significant. We estimated the individual-level SD to be about .88 point, and consequently the latter gain translates into an effect size of about .23. The average gains are also positive at 8th and 10th grade, but much smaller. Greene also found an effect for writing, but estimated it to have an effect size of 2.23.

Table 5
Average Residual Gains for FCAT Writing

GRADE	TYPE	Mean	N	SD
4	A	.04	121	.19
	B	.03	212	.22
	C	-.02	694	.22
	D	-.01	454	.24
	F	.20	66	.25
8	A	.05	73	.17
	B	.07	90	.18
	C	.00	255	.17
	D	-.05	94	.21
	F	.11	7	.17
10	A	.11	8	.09
	B	.15	12	.15
	C	.01	279	.23
	D	-.03	57	.22
	F	.10	4	.18

Greene attempted to control for regression effects by comparing higher-scoring "F" schools to lower-scoring "D" schools. "These gains made by the higher-scoring F schools in excess of what were produced by the lower-scoring D schools are what we can reasonably estimate as the effect of the unique motivation that vouchers posed to those schools with the F designation" (p. 8). Using residual scores, we repeated this analysis using 40 schools in each of the above categories aggregated across grade for reading and mathematic (though we don't suggest this as an analytic strategy). The estimates of effect were small and nonsignificant.

Discussion

The A-Plus accountability system in Florida, with its inclusion of school vouchers as one possible repercussion for low-performing schools, is a significant policy shift in the use of high-stakes assessment. Findings from evaluations of this program may thus play an important role in policy making in other states and at the federal level. Unfortunately, the Greene evaluation does not meet the methodological demands for such an evaluation. It is clear that Greene's analysis failed to account for both regression to the mean and obtaining a unique net effect of being labeled an "F" school. Sample selection is a debatable issue, and we have argued in this report that indicators based on all curriculum groups better satisfy the demands of evaluation.

Some have argued that information and research must be central to the improvement of schools:

> Schools that consistently fail to educate poor children should not receive federal dollars—and states should be accountable to Washington for ensuring that this does not happen. Federal programs that can't demonstrate results should themselves be replaced by different strategies. Though innovation and experimentation should always be encouraged, rigorous evaluation is vital and federal funds should not flow to activities that do not yield results for children. (Finn, Bruno & Ravitch, 2000)

In reply, we would argue that it's not always easy to demonstrate results given the kinds of data and accountability models that are readily available. As seen in Florida, the accountability model itself may cause some difficulty (Note 4). If schools in the lowest classification "F" improve, and yet this "improvement" is a regression artifact, then

teachers and principals and others may seize upon wholly irrelevant events as the causes of this improvement. Likewise, "D" schools that move down to the "F" classification may seize upon wholly irrelevant causes for their demise. While it is true that true "F" schools will tend to bounce up and down, and thus be more likely to become eligible for intervention, it is also true that the accountability system as currently structured may provide them with unreliable signs of their progress (or lack thereof).

Positive results are more helpful if they can be shown (by means of high quality evaluations) to be internally consistent with policy mechanisms that presumably stimulated change. One can learn better from negative outcomes if it can be shown in some detail how the policy levers failed. In other words, learning more about *how* schools made improvements or reasons for slippage is important, as well as is having confidence that the measures of loss or gain are both reliable and valid. Tying accountability to a single (or even a few) achievement outcomes has several downsides: (1) it does not automatically increase our knowledge about why things happened the way they did; (2) the use of statistical models for monitoring policy outcomes is technically demanding and requires obscure policy tools such as adjustments for regression to the mean. Moreover, it is problematic to conflate evaluation and accountability: program evaluation is intrinsically important to the mission of schools and should not be equated with establishing "results" as defined by Washington.

We can agree that hard-nosed evaluation is necessary, but it is useful to expand on what such evaluation activities should include:

Technical considerations. The state of Florida should consider methods that are used elsewhere (e.g., North Carolina) to stabilize the indicators that are used to designate school classifications. Such models use past achievement data to estimate expected growth, and designate exemplary growth in a manner that controls for some statistical artifacts such as regression. Though there are costs associated with a more complex model, the decision to focus accountability on test scores requires more sophisticated statistical apparatus within the accountability model. Moreover, focus on a small set of indicators accompanied by significant sanctions can force schools to employ instructional methods that are optimized for short-term payoffs. Consequently, additional accountability components may be required to monitor for negative consequences such as an increase in the number of remedial classes, focusing on test preparation, curricular materials that are substantially similar to test preparation material, and increases in drop-out rates.

Policy considerations. One of the most important roles of policy evaluation is to inform policymakers not only about whether or not a program is working, but *why* it is having the noted effects. Evaluations that provide little or no information about the mechanisms that have led to reported changes are both less compelling and more subject to criticism. In Florida, there is currently little information about what schools are doing that would lead one to expect that scores would improve. This information is crucial for the future development of the accountability program and might include, for example, an evaluation of capacity within schools identified as needing intervention, or an analysis of how administrative rules are interpreted by local staff. Policy makers should also receive evaluation information regarding the accountability model or system itself as well as behavior that is the object of the model.

In the case of Florida, this report suggests that it is simply not clear whether or not the threat of vouchers is having a positive impact on student test scores. There is some evidence of a small effect at 8th grade in reading and mathematics, and in writing at 4th grade. These findings should be investigated in a more thorough analysis (taking into account, for example, exclusion rates). If these findings withstand further analysis, it

would also be important to examine a number of potential causes including resources (e.g., professional development or teaching materials), school intervention plans, staffing changes, and other taken remedies to improve performance. In other words, it is overly simplistic to assume that the voucher threat was the only active agent, or that other causes were contingent on the voucher threat.

Conclusion

We offer an alternative to Greene's generous and simplistic reading of the evidence. At face value, the large gains (as seen in effect sizes of .80, 1.25, and 2.23 for reading, mathematics, and writing) were implausible and should have been submitted to additional methodological scrutiny. Upon such an examination, we have raised serious questions regarding the validity of Greene's empirical results and conclusions. Indeed, one *should* follow the general advice to expect horses when one hears hoof beats, not unicorns.

Notes

1. These two schools were chosen in 1999 for the voucher plan in the first year of the accountability policy implemented in 1999. These schools did not meet the "2-out-of-4" policy, but had received an "F" in 1999 and appeared on a 1998 list of low performing schools (Sandham, 1999).
2. It could be argued that the group of students who were used to determine the school grade might also be the appropriate sample. It appears that "standard curriculum" designates eligible students. According to the State Board of Education Administrative Rules (6A-1.09981)

   > (3)(a) For the purpose of calculating state and district results, the scores of all students enrolled in standard curriculum courses shall be included. This includes the scores of students who are speech impaired, gifted, hospital homebound, and Limited English Proficient (LEP) students who have been in an English for Speakers of Other Languages (ESOL) program for more than two (2) years.

   To receive a grade of "D" or higher, schools are required to test at least 90% of their eligible students. There are additional restrictions on student inclusion for determining school grade in 6A-1.09981:

   > (3)(b) For the purpose of designating a school's performance grade, only the scores of those students used in calculating state and district results who are enrolled in the second period and the third period full-time equivalent student membership survey as specified in Rule 6A-1.0451, FAC., shall be included.

   Because these criteria, fairly applied, may create inconsistencies across schools, the group of all students tested may provide a school average better for the purposes of evaluation. It would also be useful to have the school median and exclusion rates.
3. The frequencies in Tables 3 and 4 differ slightly from the actual number of schools in each category. For example, 5 high schools received grades of "F," yet

there are only 4 in our study. In checking this result, we found that the 5th high school was no longer listed in official documents in 1999-2000. Other than this difference, however, our data agree with state data in terms of the numbers of "F" schools for elementary, middle, and high schools.

4. We note that only two of the schools on the 1998 list of critically low performing schools received an "F" in 1999. Likewise none of the 78 schools receiving an "F" in 1999 also received an "F" in 2000; however, only 4 schools received an "F" in 2000.

References

Campbell, D.T., and Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Finn, C. E., Manno, B. V., Jr. & Ravitch, D. (2000). Education 2001: Getting the Job Done A Memorandum to the President-Elect and the 107th Congress. Thomas B. Fordham Foundation, Washington, D.C.: author.

Florida Department of Education. (2001). *FCAT Briefing Book*, Florida Department of Education, Tallahassee: author.

Glass, G. V & Hopkins, K.D. (1996). *Statistical Methods in Education and Psychology (3rd Ed.)*. Boston: Allyn and Bacon.

Glass, G.V & Robbins, M.P. (1967). A critique of experiments on the role of neurological organization in reading performance. *Reading Research Quarterly*, 3, 5-51.

Greene, J. P. (2001a). *An Evaluation of the Florida A-Plus Accountability and School Choice Program*. New York: The Manhattan Institute.

Greene, J. P. (2001b). Bush's School Plan: Why We Know It'll Work. *New York Post* (February 21).

Greene, J. P., Peterson, P. E., & Du, J. (1998). School Choice in Milwaukee: A Randomized Experiment. In P. E. Peterson & B. C. Hassel (Eds.), *Learning from School Choice*. Washington, D.C.: Brookings Institution Press.

Hedges, L.V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F. & Stecher, B. M. (2000). What Do Test Scores in Texas Tell Us? *Education Policy Analysis Archives,* 8 (49).

Lopez, K. J. (2001). Bush Ed Plan Rates an "A." *National Review Online* (February 22).

Public Schools of North Carolina. (2000). Setting Annual Growth Standards: "The Formula." North Carolina Department of Public Instruction, North Carolina: author.

Sandham, J. L. (1999). Schools Hit by Vouchers Fight Back. *Education Week* (September 15).

Schemo, D. J. (2001). Threat of Vouchers Motivates Schools to Improve, Study Says. *New York Times* (February 16).

Witte, J. F. (1996). Who Benefits from the Milwaukee Choice Program? In B. Fuller & R. F. Elmore (Eds.), *Who Chooses? Who Loses?* New York: Teachers College Press.

About the Authors

Gregory Camilli
Professor
Rutgers Graduate School of Education

Email: camilli@rci.rutgers.edu

Gregory Camilli is Professor of Educational Psychology, at the Rutgers Graduate School of Education and a Senior Researcher in the Center for Educational Policy Analysis. His areas of research interest include psychometric issues in educational policy, meta-analysis, and differential item functioning. Examples of recent publications include "Values and state ratings: An examination of the state-by-state education indicators in quality counts" (*Educational Measurement: Issues and Practice*, 2000), "Application of a method of estimating DIF for polytomous test items" (*Journal of Educational and Behavioral Statistics,* 1999), Standard error in educational programs: A policy analysis perspective" (*Educational Policy Analysis Archives*, 1996), and *Methods for Identifying Biased Items* (Sage, 1994). Camilli has been or is currently a member of the editorial Boards of *Educational Measurement: Issues and Practice*, *Educational Policy Analysis Archives*, and *Education Review*. He is a regular reviewer for *Applied Measurement in Education, Journal of Educational Measurement, Psychometrika*, and *Psychological Methods*, among others. As a member of the Technical Advisory Committee of the New Jersey Basic Skills Assessment Council, he provides expertise on testing and measurement issues to the New Jersey state assessment program.

Katrina Bulkley
Assistant Professor
Rutgers Graduate School of Education

Email:bulkley@rci.rutgers.edu

Katrina Bulkley is an Assistant Professor of Educational Policy at the Rutgers University Graduate School of Education. Much of her work has focused on issues involving school choice and charter schools. Recent articles include, "Charter School Authorizers: A New Governance Mechanism?" in *Educational Policy* (November 1999), and "'New Improved' Mayors Take Over City Schools" (with Michael Kirst) in Phi Delta *Kappan* (March 2000). She is currently working with the Consortium for Policy Research in Education on a literature review of research on charter schools and a study of for-profit management companies and charter schools, and with the Center for Education Policy Analysis, located at Rutgers University, on two studies of the impact of standards, testing and professional development on instructional practices in New Jersey. Bulkley has reviewed articles for *Educational Evaluation and Policy Analysis*, *Educational Policy*, and *Policy Studies Journal*.

The World Wide Web address for the *Education Policy Analysis Archives* is **epaa.asu.edu**

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-0211. (602-965-9644). The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

EPAA Editorial Board