Education Policy Analysis Archives

Volume 9 Number 8 **March 19, 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*.

The Effects of Vouchers on School Improvement: Another Look at the Florida Data

Haggai Kupermintz
University of Colorado at Boulder

Related article.

Abstract
This report re-analyzes test score data from Florida public schools. In response to a recent report from the Manhattan Institute, it offers a different perspective and an alternative explanation for the pattern of test score improvements among low scoring schools in Florida.

Introduction

A recent report from the Manhattan Institute think tank (Greene, 2001) examined test scores of Florida public schools in 1999 and 2000 to determine the effects of vouchers on student performance. The report ends with a conclusion: “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.” My own analyses of the Florida data lead to no such conclusion. Instead, I found the evidence telling a more interesting, and to my mind a more believable, story. I will argue that the evidence suggests that the “voucher effect” follows different patterns in the three tested subject areas: reading, math, and writing. Moreover, I will show that the most dramatic improvements in failing schools were realized by targeting and achieving a minimum “passing” score on the writing test, thereby escaping the threat of losing their students to vouchers.

Background

The Florida A-Plus school accountability program is based on tracking schools' performance and progress toward the educational goals set in the Sunshine State Standards. The main source of information on school performance is a series of standardized test in reading, math, and writing, known collectively by the somewhat redundant name FCAT (Florida Comprehensive Assessment Tests). All elementary, middle, and high school students are tested annually (different subjects in different grades) and the results are used to assign a grade to each school, from A to F, according to a formula that weighs the number of students performing below and above pre-defined markers along the test score scales. An F grade assignment has a variety of consequences and a great deal of attention is directed toward F schools in the Florida system.

One of the most visible and politically contested consequences of failing the State's tests is the voucher provision. If a school received another F grade in a four-year period, its students become eligible to take their public funding elsewhere to a private or better-performing public school. In 1999, 78 schools have received an F grade. Greene's report examines the gains these schools made on the FCAT between 1999 and 2000, and the executive summary offers a précis of the evidence: “The results show that schools receiving a failing grade…achieved test score gains more than twice as large as those achieved by other schools. While schools with lower previous test scores across all state-assigned grades improved their test scores, schools with failing grades that faced the prospects of vouchers exhibited especially large gains” (Greene, 2001, p. ii). The report itself compares the average score gains of higher-scoring F schools to lower-scoring D schools, serving as a control group. Standardized group differences constitute Greene's estimated effect sizes of the “voucher effect”—0.12 in reading, 0.30 in math, and 0.41 in writing. Other analyses in the report calculate the correlations between FCAT and other standardized test administered in Florida schools, to gauge the validity of the FCAT.

These findings lead Greene not only to the conclusions cited above, but also to strong public commentary in the local and national press in favor of Florida's voucher system and similar proposals in President Bush's school reform plan. The moderate “voucher effect” estimates and relatively cautious language of the report were replaced in the media by strong statements, emphasizing the magnitude of the raw score gains achieved by F schools. In an interview to the St. Petersburg Times (February 16, 2001), after the release of his report, Greene asserted: "The F schools showed tremendous gains because they faced a particularly concrete outcome that they wished to avoid: embarrassment, loss of revenue, vouchers”. Even more boldly, generalizing from the Florida findings, Greene offered the following proclamation in a guest commentary in The New York Post (February 21, 2001): “So the improvement by Florida's failing schools was real. So, as debate proceeds over President Bush's education proposals, know this: Testing, accountability and choice are powerful tools to improve education - and, in particular, to turn around chronically failing schools. That's not a theory, but proven fact.”

My re-analyses of the Florida data suggest that Greene might have over-stated the case for the simple explanation he promoted in his report and in the press. A more careful examination of the patterns of gains reveals that failing schools responded with a more sophisticated strategy than the undifferentiated, gross “voucher effect” gave them credit for. The key element of the strategy was to achieve a particular score on the writing test, in order to elevate their grades. The strategy was extremely successful and all failing schools were able to escape the threat of vouchers by achieving a grade of D or better in 2000.

Data

The data for the analyses are school mean scores on the FCAT reading, math, and writing tests from 1999 and 2000. They include all curriculum groups in both years (available on-line from the Florida Department of Education web site: http://www.firn.edu/doe/sas/fcat.htm). These data are slightly different from the data Greene used in his analyses, but as he comments (Greene, 2001, Note 10), the difference is inconsequential and similar conclusions will be reached using either dataset. The analyses below address issues that Greene either paid no attention to in his report or dismissed as unimportant. The first example of the latter is regression toward the mean.

An elusive regression artifact

On page 10 of his report, Greene alerts his readers to the potential biasing affect of regression to the mean:

> As another alternative explanation critics might suggest that F schools experienced larger improvements in FCAT scores because of a phenomenon known as regression to the mean. There may be a statistical tendency of very high and very low-scoring schools to report future scores that return to being closer to the average for the whole population. This tendency is created by non-random error in the test scores, which can be especially problematic when scores are "bumping" against the top or bottom of the scale for measuring results. If a school has a score of 2 on a scale from 0 to 100, it is hard for students to do worse by chance but easier for them to do better by chance. Low-scoring schools that are near the bottom of the scale are very likely to improve, even if it is only a statistical fluke.

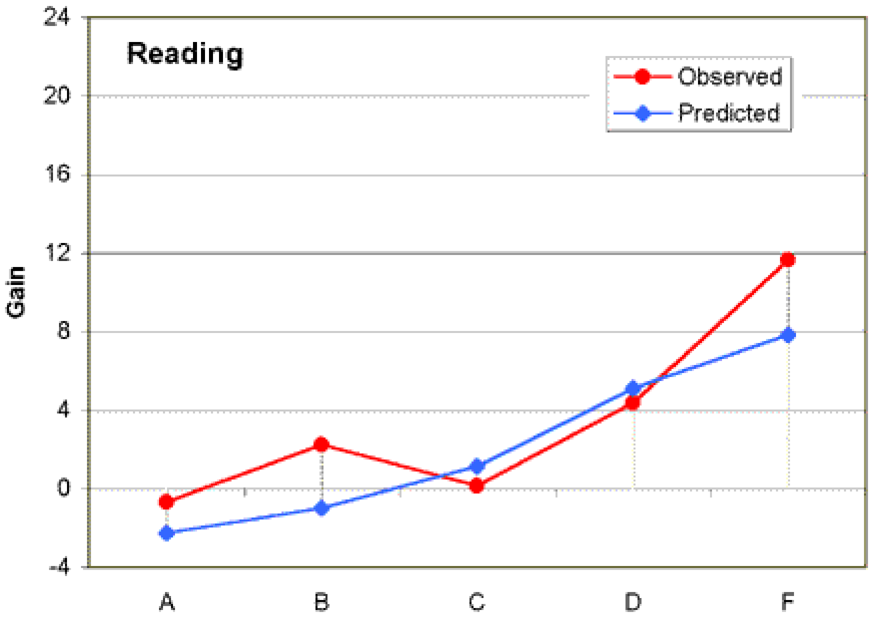
He then dismisses the threat because "the scores of those [F] schools were nowhere near the bottoms of the scale of possible scores" (p. 10). Greene seems to confuse regression toward the mean with floor and ceiling effects–completely different phenomena. Scores "'bumping' against the top or bottom of the scale" colorfully characterizes ceiling and floor effects but is an inadequate description of the regression effect. Regression toward the mean operates whenever the correlation between two variables (the 1999 and 2000 test scores, in our case) is less than perfect. It influences the entire range of scores—not just the very extreme—with a force proportional to their distance from the sample mean. Therefore, the fact that F schools where far from the bottom of the score scale is a poor indication that regression effects are absent. The two relevant pieces of information are how far the group is *from the sample mean* and the magnitude of the correlation between the two variables involved. Knowing these two quantities allows us to forecast the expected magnitude of the pull toward the sample mean. Using standardize scores aids interpretation, as the predicted standardized Y equals $Zy = rZx$ (X and Y are the 1999 and 2000 test scores, respectively). For example, a school 2 standard deviation below the mean in 1999 will be expected to score only .85(2) = 1.7 standard deviations below the mean in 2000, assuming a correlation of .85 (a value compatible with the typical correlation is the Florida data)—an effect size of .3! In 1999, F schools were 1.9*SD*s below the mean in reading, 1.7*SD*s below the mean in math, and 1.8*SD*s below the mean in writing. This simple analysis shows that the excepted magnitude of the regression effect warrants serious attention.

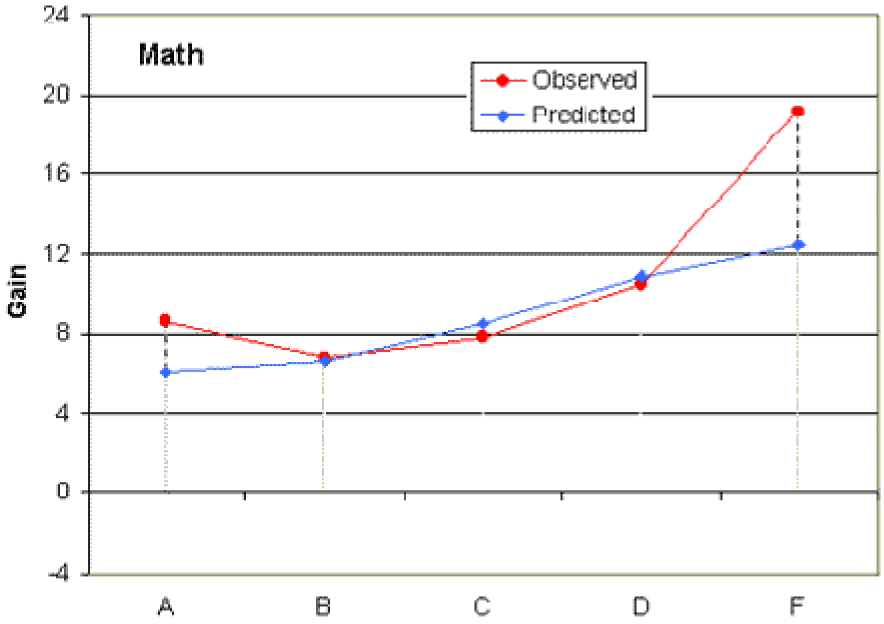
Using a slightly more complicated formula (see, e.g., Campbell & Kenny, 1999, p. 28, Table 2.1), and the regression coefficient instead of the correlation, one can calculate the expected 2000 score or the expected score gain, given a particular level of performance in 1999. Table 1 gives the expected score gains, if regression toward the mean was the only factor responsible for these gains, for the three FCAT tests, alongside with the observed gains for schools with different grades in 1999 [Note 1]. Figure 1 shows the same findings graphically.

Table 1
Predicted and Observed Gains By School Grade

	Reading		**Math**		**Writing**	
Grade	Observed	Predicted	Observed	Predicted	Observed	Predicted
A	-.68	-2.29	8.62	6.11	.24	.27
B	2.24	-1.01	6.85	6.65	.27	.29
C	.15	1.13	7.83	8.47	.29	.30
D	4.37	5.12	10.47	10.90	.33	.33
F	11.64	7.81	19.18	12.42	.67	.37

Figure 1. Predicted and Observed Gains By School Grade


Reading
Observed
Predicted
Gain
24
20
16
12
8
4
0
-4
A
B
C
D
F


Math
Observed
Predicted
Gain
24
20
16
12
8
4
0
-4
A
B
C
D
F

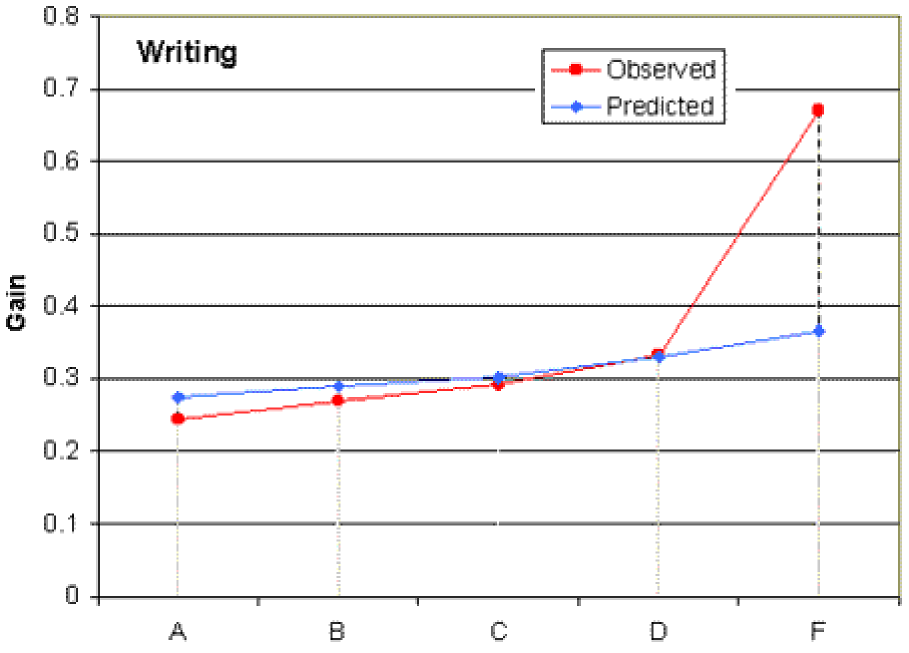
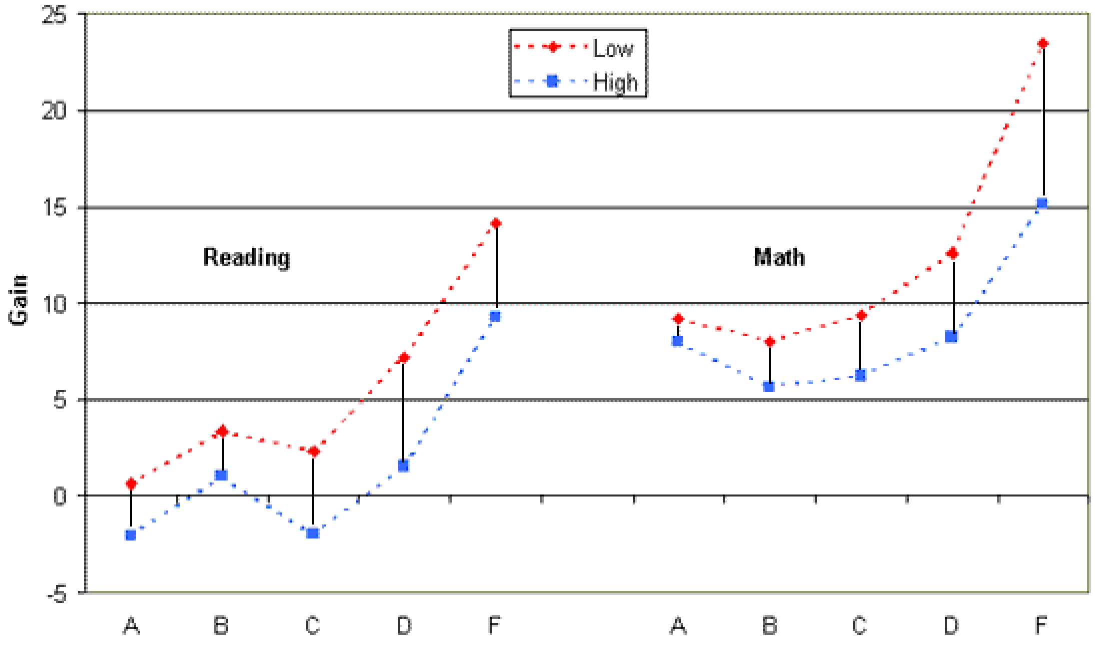


Figure 1 portrays an interesting picture. The height of each red dot represents the observed gain in scores between the 1999 and 2000 administrations of the FCAT. The blue dots represent the predicted gains attributed to the regression effect, and the distance between the red and blue dots, connected by a dashed line, depicts the "residual gain"—the amount of gain left after the regression effect has been accounted for. From Figure 1 we learn that a substantial portion (67% in reading, 64% in math, and 55% in writing [Note 2]) of the observed gains among F schools is due to regression to the mean. Note also that F schools do not appear exceptional and their residual gains are comparable to those observed in B schools, for example. These schools, however, start to stand out when we examine the patterns in math and even more so in writing. These observations agree with the order of effect sizes reported by Greene in Table 3 of his report. Unfortunately, Greene stops here to conclude: "a voucher effect." But the story has just begun to unfold.

Within-group patterns

We now direct our attention to the patterns of change within each group of schools designated by the same grade. In his second response to the potential regression threat, Greene suggested that "if the improvements made by f schools were concentrated among those F schools with the lowest previous scores, then we might worry that the improvements were more of an indication of regression to the mean (or bouncing against the bottom) than an indication of the desire to avoid having vouchers offered in failing schools". Curiously, while Greene argues for this strategy he never conducts the analysis. Instead, he presents in Table 5 *residual gains* that already take the regression effect into account. Even then he ignores the large difference between lower and higher scoring F schools in writing. Ironically, this difference is 0.16, exactly equal to the "voucher effect" in writing! Moreover, the same rationale for using residual gains here should apply with equal force for the gains reported elsewhere in Greene's report. The basic logic remains the same between tables.

Figure 2. Observed Gains by Initial Status and School Grade



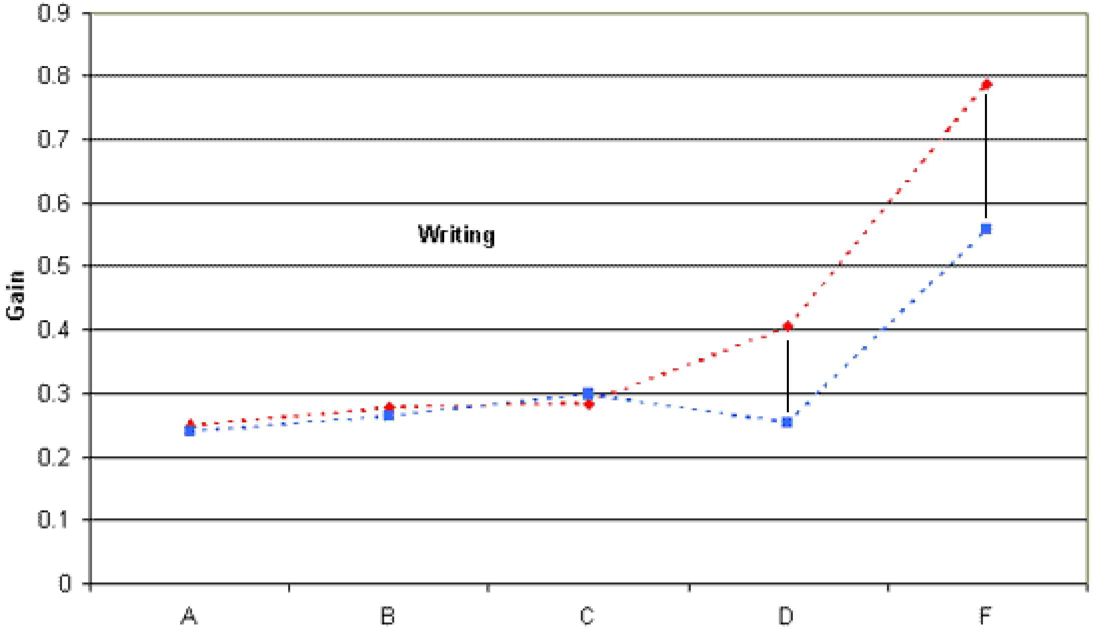


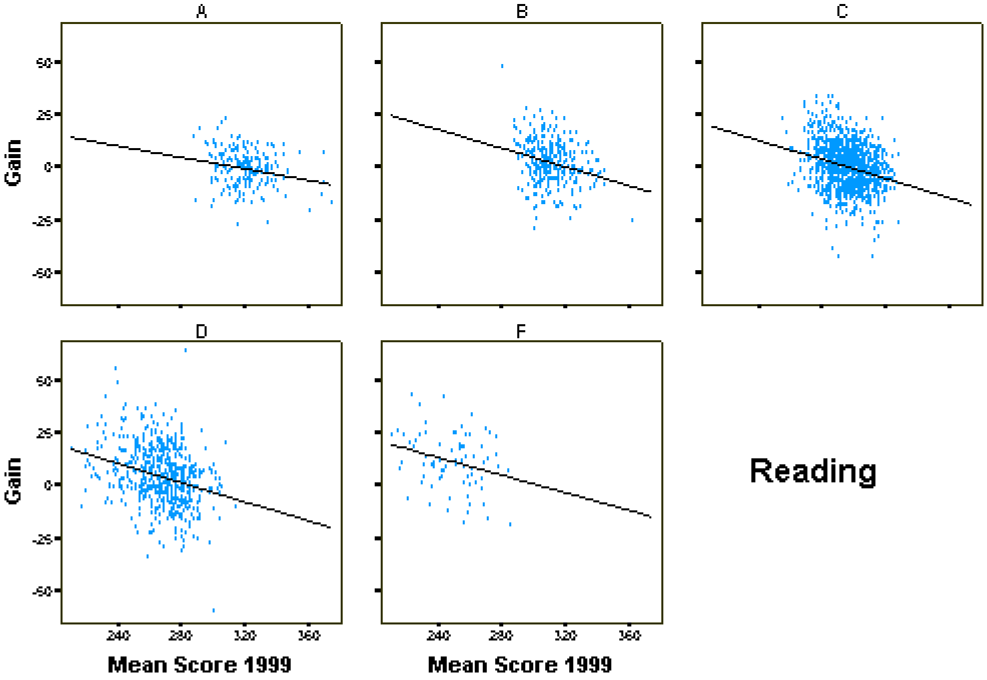
Figure 2 might cause us to worry, as Greene was right to point out. The red dots are the average gains made by the lower scoring schools (below the group median [Note 3]) and the blue dots the average gains made by higher scoring schools (above the group median) in each grade group. While the differences between gains of lower and higher scoring schools are constant across grade groups for reading, they increase substantially as grades get lower for math. For writing only, D and F schools show within-group differences, and these are more pronounced among F schools. In fact, the difference between higher and lower scoring F schools in writing is 0.23 representing an effect size of 0.23/0.39 = 0.6, substantially larger than the largest voucher effect Greene reports (an effect size of 0.41 in writing, see Table 3 in Greene's report)!

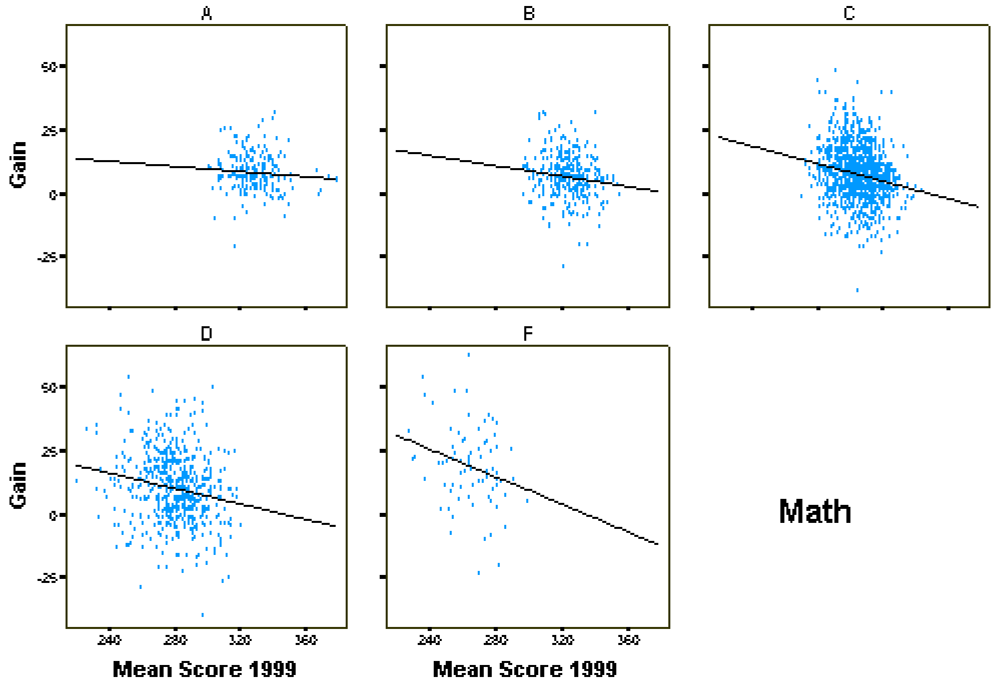
The within group analysis needs to be refined further as we change lens to zoom in on the details of patterns of gains within the different grade groups. Figure 3 shows the scatter plots of the 1999 and 2000 scores with the linear fits superimposed and depicting the overall trends in the data. Table 2 complements the graphs by giving the standardized regression coefficients corresponding to the trend lines.

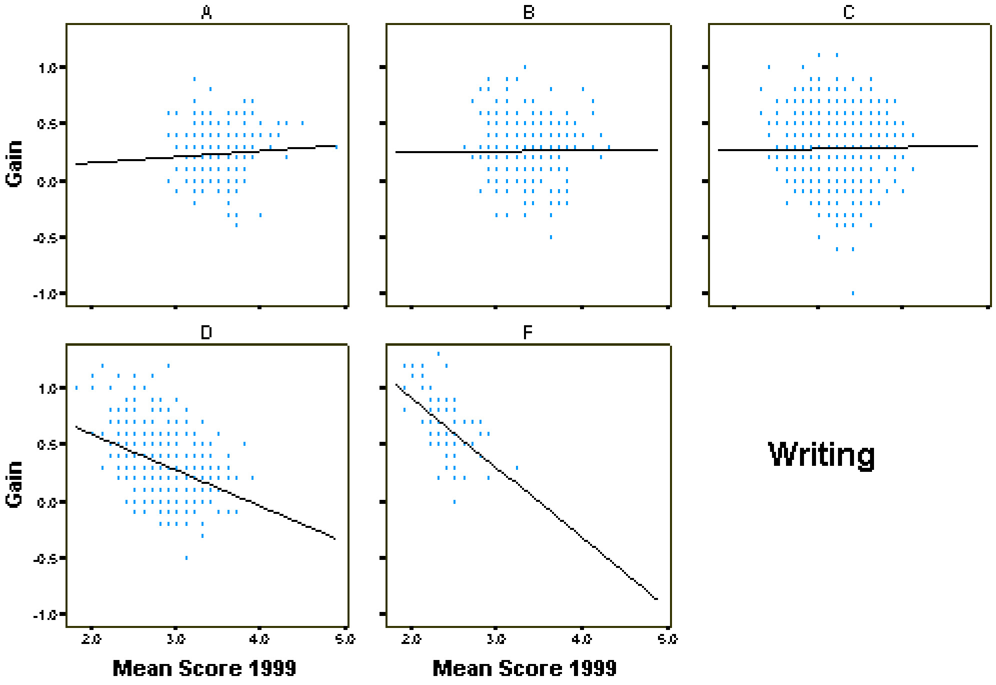
Table 2
Standardized Regression Coefficients of Gains Predicted from 1999 Scores

Grade	Reading	Math	Writing
A	-0.23	-0.09	0.07
B	-0.26	-0.14	0.01
C	-0.27	-0.20	0.02
D	-0.28	-0.19	-0.39
F	-0.28	-0.26	-0.54

Figure 3. Gains as a Function of 1999 Scores by School Grade





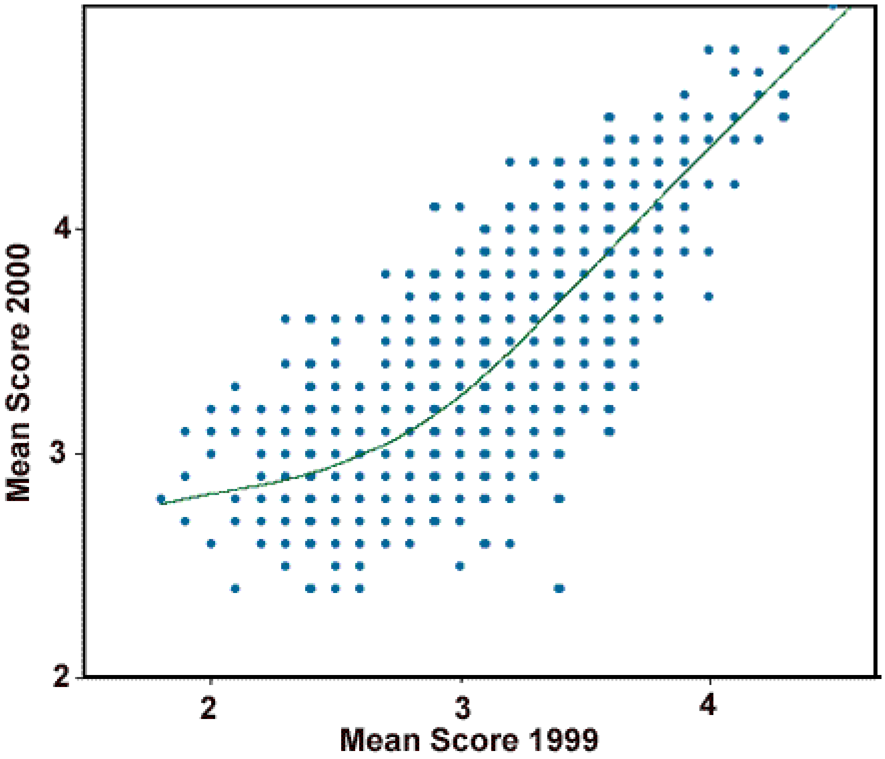


The reading scores behave as expected—a moderate negative correlation in all grade groups between the score achieved in 1999 and the gain realized one year later. Consistent with the patterns we identified in the cruder comparisons of Figure 2, the link between prior scores and gains becomes stronger as grades go down, a pattern most pronounced in writing. The findings for writing are striking. The amount of gain in F schools, and to a lesser extent D schools, is strongly determined by how low their scores were in 1999; the standardized regression coefficient is -0.54, representing the effect size of the mean gain difference for schools that scored one standard deviation apart from each other in 1999 (closely resembling the effect size value for lower and higher scoring F schools we calculated before). This pattern is completely absent for A, B, and C schools, whose 1999 scores provide no information on their expected gain.

The writing on the wall

The seemingly curious pattern of gains for writing has, in fact, a simple explanation. If there was a clear mark on the writing score scale that D and F schools set up to reach, not more nor less, then lower scoring schools would have to close a wider gap to reach the mark, giving rise to a strong negative correlation between where they started and how far they had to go (their gain). Figure 4 clearly demonstrates this phenomenon. It shows, for the entire school population, the relationships between 1999 scores and 2000 mean scores and gains. The lines represent the best fitted nonlinear trend lines (using the "loess" technique, see Chambers & Hastie, 1991, pp. 309-376).

Figure 4. Writing 2000 Scores and Gains as a Function of 1999 Scores



References

Campbell, D. T., & Kenny, D. A. (1999). *A primer of regression artifacts*. New York: Guilford Press.

Chambers, J. M. and T. J. Hastie, Eds. (1991). *Statistical models in S*. Pacific Grove, CA: Wadsworth & Brooks/Cole.

Cronbach, L.J. and Associates. (1980). *Toward reform of program evaluation.* San Francisco CA: Jossey-Bass

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

About the Author

Haggai Kupermintz
School of Education
University of Colorado at Boulder

Email: haggai.kupermintz@colorado.edu

Haggai Kupermintz is an Assistant Professor of research and evaluation methodology at the University Colorado at Boulder, School of Education. His specializations are educational measurement, statistics, and research methodology. His current work examines the structure, implementation, and effects of large-scale educational accountability systems.

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

EPAA Spanish Language Editorial Board