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

CSE Technical Report 558

Haggai Kupermintz
CRESST / University of Colorado at Boulder

April 2002

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 1.1. Comparative Analyses of Current Assessment and Accountability Systems
Robert L. Linn, Project Director, CRESST/University of Colorado at Boulder

The work reported herein was supported in part under the Educational Research and Development Centers Program, PR/Award Number R305B960002-01, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

THE EFFECTS OF VOUCHERS ON SCHOOL IMPROVEMENT: ANOTHER LOOK AT THE FLORIDA DATA[1]

Haggai Kupermintz
CRESST/University of Colorado at Boulder

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.

A recent report from the Manhattan Institute (Greene, 2001a) examined test scores of Florida public schools in 1999 and 2000 to determine the effects of vouchers on student performance. The report ended 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" (p. 11). My analyses of the Florida data did not lead to this conclusion. Instead, I found the evidence telling a more interesting, and to my mind a more believable, story. In this report I argue that the evidence suggests that the "voucher effect" follows different patterns in the three tested subject areas: reading, math, and writing. Moreover, I 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 tests in reading, math, and writing, known collectively as the FCAT (Florida Comprehensive Assessment Tests). All elementary, middle, and high school

[1] My thanks go to Greg Camilli, Sherman Dorn, Steve Lang, Bob Linn, Lorrie Shepard, and Kevin Welner for helpful comments.

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 receives a second F grade in a 4-year period, its students become eligible to take their public funding elsewhere to a private or better performing public school. In 1999, 78 schools received an F grade. Greene's (2001a) 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" (p. ii). The report 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 the FCAT and other standardized tests administered in Florida schools, to gauge the validity of the FCAT.

My re-analyses of the Florida data suggest that Greene might have overstated the case for the simple explanation he offered in his report. 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 online 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, 2001a, Note 10), the difference is inconsequential, and similar conclusions will be reached using either dataset. The analyses below address issues that Greene either did not discuss in his report or regarded as not significant. The first example is regression toward the mean.

An Elusive Regression Artifact

On page 10 of his report, Greene (2001a) alerts his readers to the potential biasing effect of regression toward 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 rejects the threat because "the scores of those [F] schools were nowhere near the bottom of the scale of possible scores" (Greene, 2001a, p. 10). Greene seems to mix regression toward the mean with floor and ceiling effects—two different phenomena. Scores "'bumping' against the top or bottom of the scale" 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 this 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 were far from the bottom of the score scale is not a strong 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 deviations below the mean in 1999 will be expected to score only $.85 \times 2 = 1.7$ standard

deviations below the mean in 2000, assuming a correlation of .85 (a value compatible with the typical correlation in the Florida data)—an effect size of .3. In 1999, F schools were 1.9 standard deviations below the mean in reading, 1.7 standard deviations below the mean in math, and 1.8 standard deviations 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, along with the observed gains for schools with different grades in 1999.[2] Figure 1 shows the same findings graphically.

Figure 1 portrays an interesting picture. The height of each red dot (square) represents the observed gain in scores between the 1999 and 2000 administrations of the FCAT. The blue dots (diamonds) represent the predicted gains attributed to the regression effect, and the distance between the red (square) and blue (diamond) 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[3]) of the observed gains among F schools is due to regression to the mean. Note also that

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

[2] The calculations of the regression coefficients in these analyses excluded F schools to avoid attributing a potential true program effect to the regression artifact.

[3] These percentages are calculated as the observed gain divided by the predicted gain and multiplied by a hundred. For example the figure for reading is (7.81 / 11.64) x 100 = 67%.

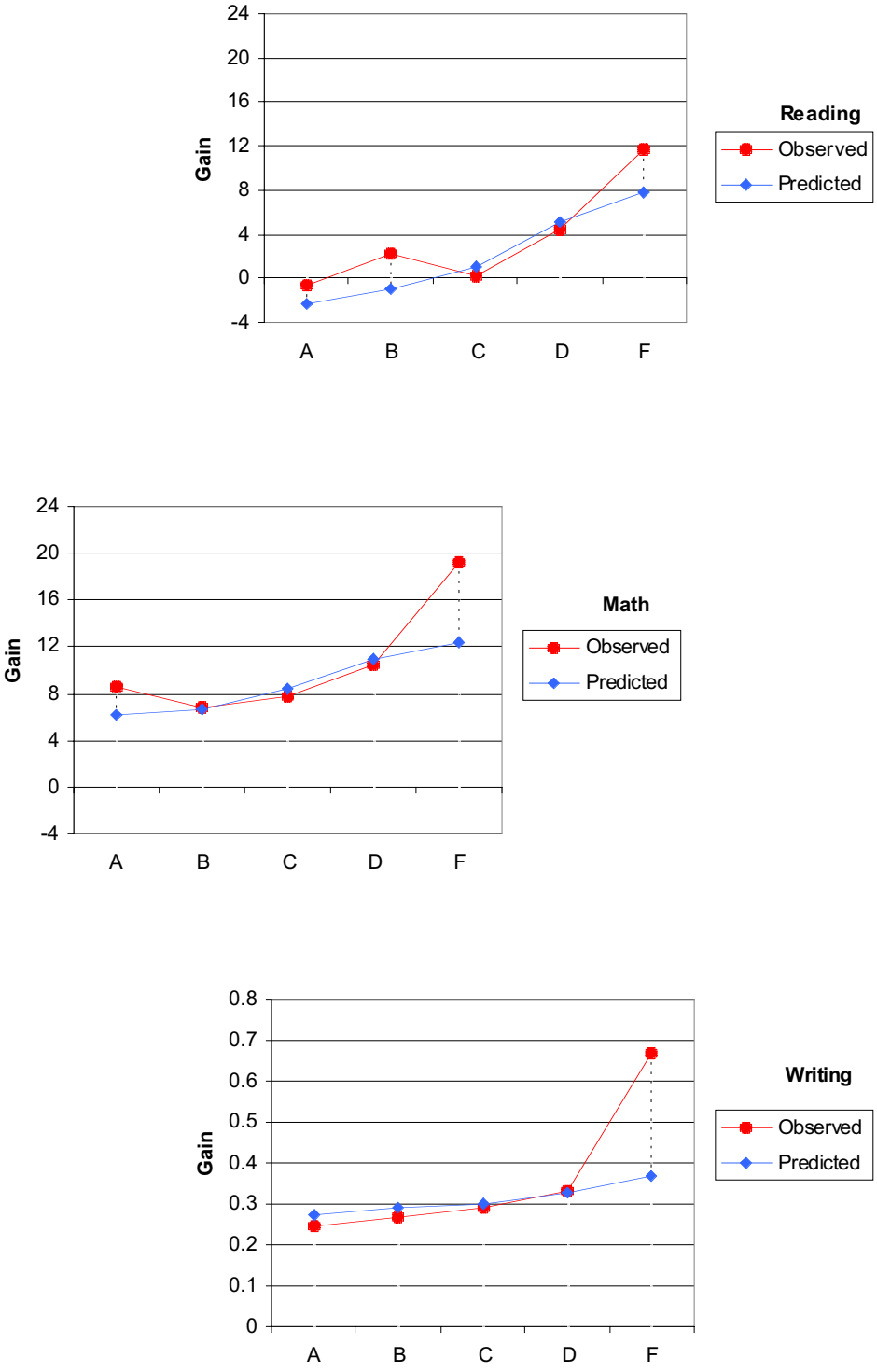

Figure 1. Predicted and observed gains by school grade.

F schools do not appear exceptional, and in reading 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 but even more so in writing. These observations agree with the order of effect sizes reported by Greene (2001a) in Table 3 of his report. Greene stopped 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 (2001a) 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" (p. 10). While Greene argued for this strategy, he never conducted the analysis. Instead he presented in Table 5 *residual gains* that already take the regression effect into account. Even then he ignored the large difference between lower and higher scoring F schools in writing. Ironically, this difference is 0.16 and exactly equals 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 might cause us to worry, as Greene was right to point out. The red dots (now diamonds) are the average gains made by the lower scoring schools (below the group median[4]) and the blue dots (now squares) 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 (2001a) reported (an effect size of 0.41 in writing; see Table 3 in Greene's report).

[4] The choice between the mean and median is inconsequential in this analysis. I used the median because it produces slightly more equal sample sizes.

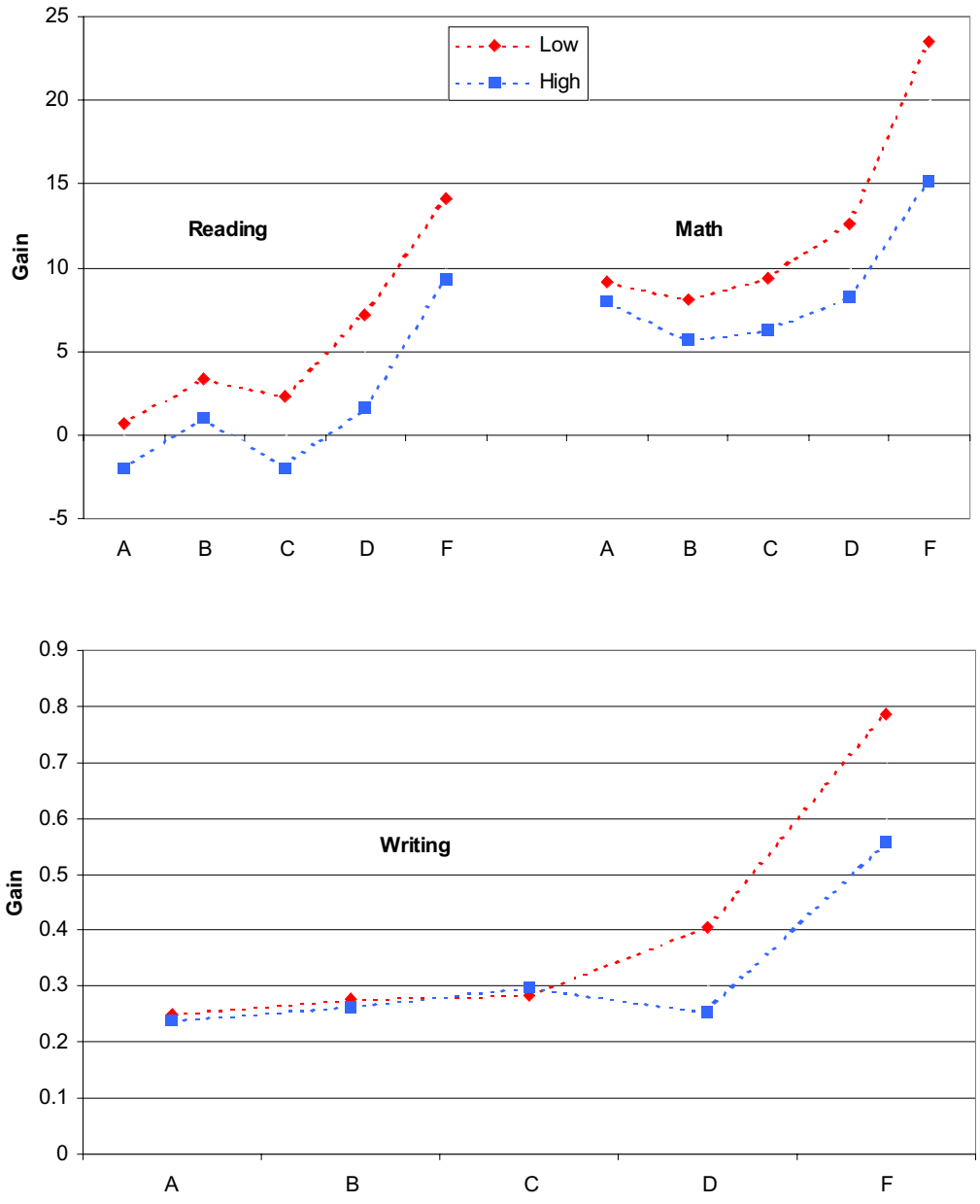

Figure 2. Observed gains by initial status and school grade.

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.

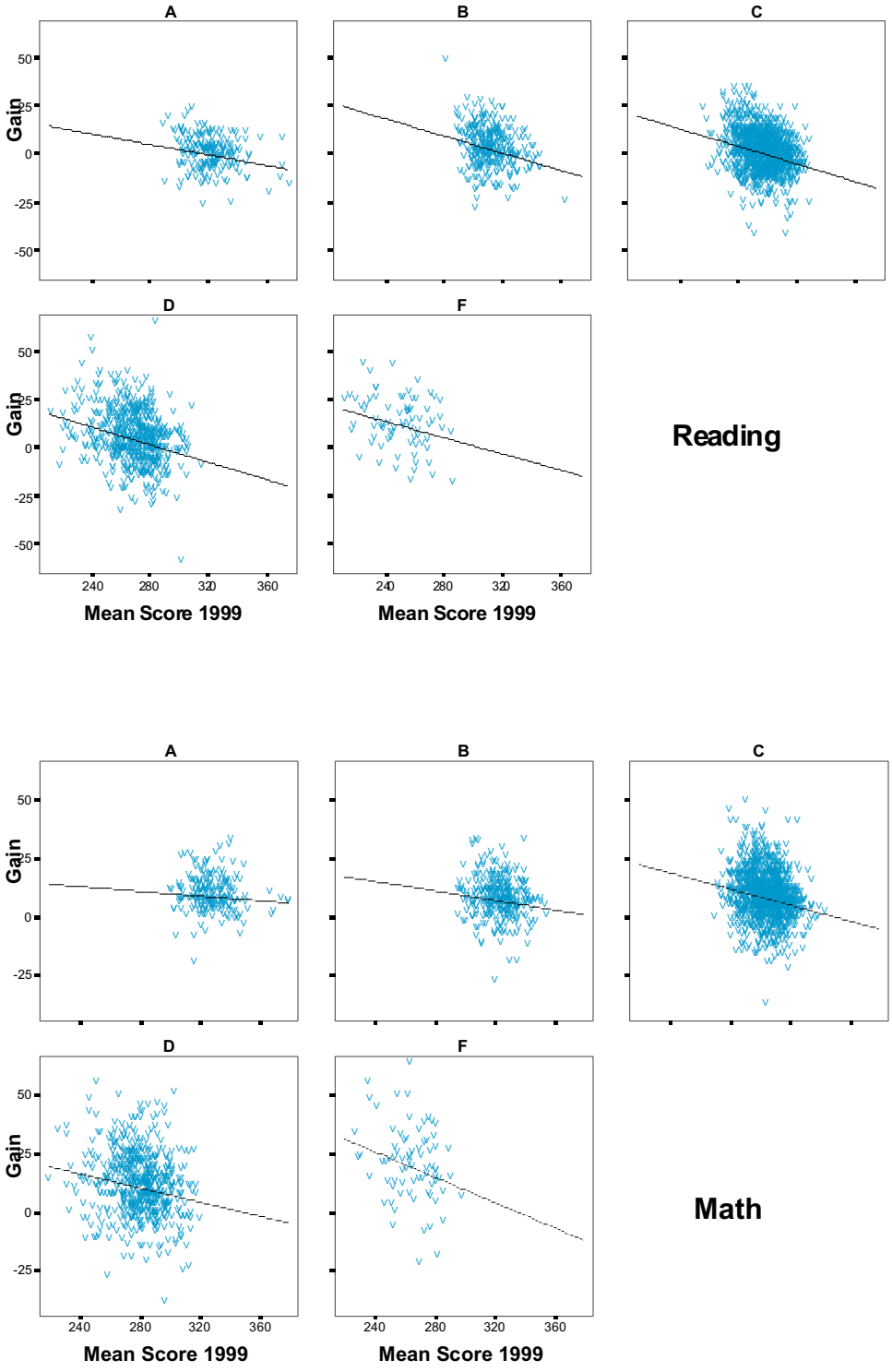

Figure 3. Gains as a function of 1999 scores by school grade.

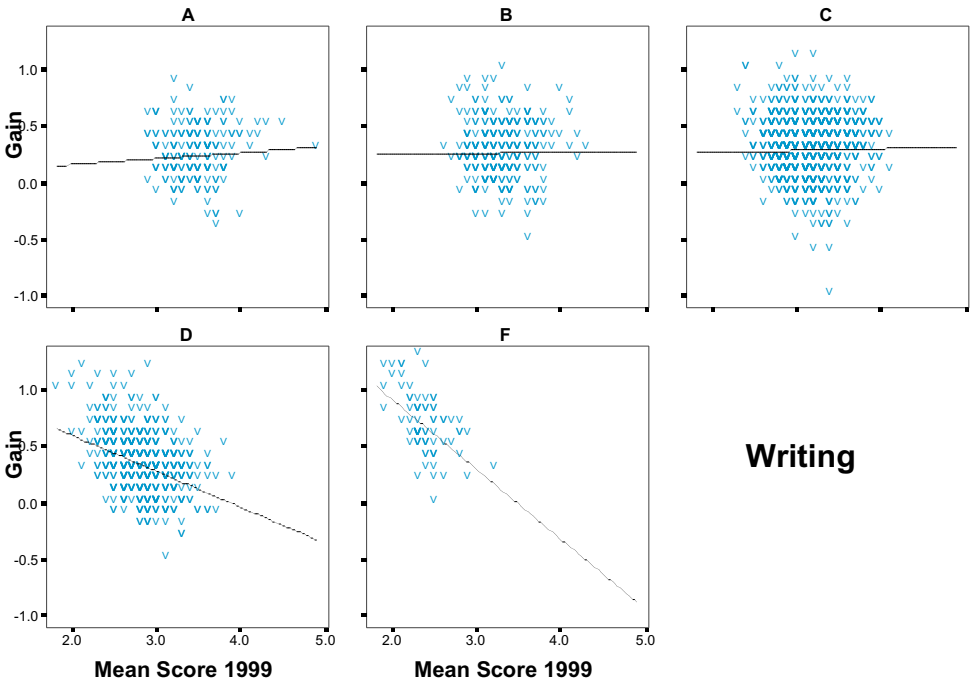

Figure 3. (continued).

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

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 in 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 the difference between 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 reading in fact has a simple explanation. If there was a clear mark on the writing score scales that D and F schools set up to reach—not more, not 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 strongly suggests that the mark was a score of 3.0 on the writing test. Schools that scored less than 3.0 in the 1999 assessment have managed to make up the difference and reach the mark in 2000. The gain slope starts an upward bend below 3.0 in 1999—schools that scored less than 3.0 in 1999 have stabilized their performance around a score of 3.0 in 2000.

Conclusion

On June 21, 2000, long before the release of the Manhattan Institute report, the *St. Petersburg Times* ran a story entitled “Why Are Florida Children Writing so Much Better?” (Hegarty, 2000). Noting the impressive improvement in the writing score, the story offered an explanation: “How could so many kids suddenly become competent writers? Many educators were not completely surprised at the improvement. Out of fear and necessity, Florida educators have figured out how the state’s writing test works and are gearing instruction toward it—with constant writing and, in many cases, a shamelessly formulaic approach. For some struggling

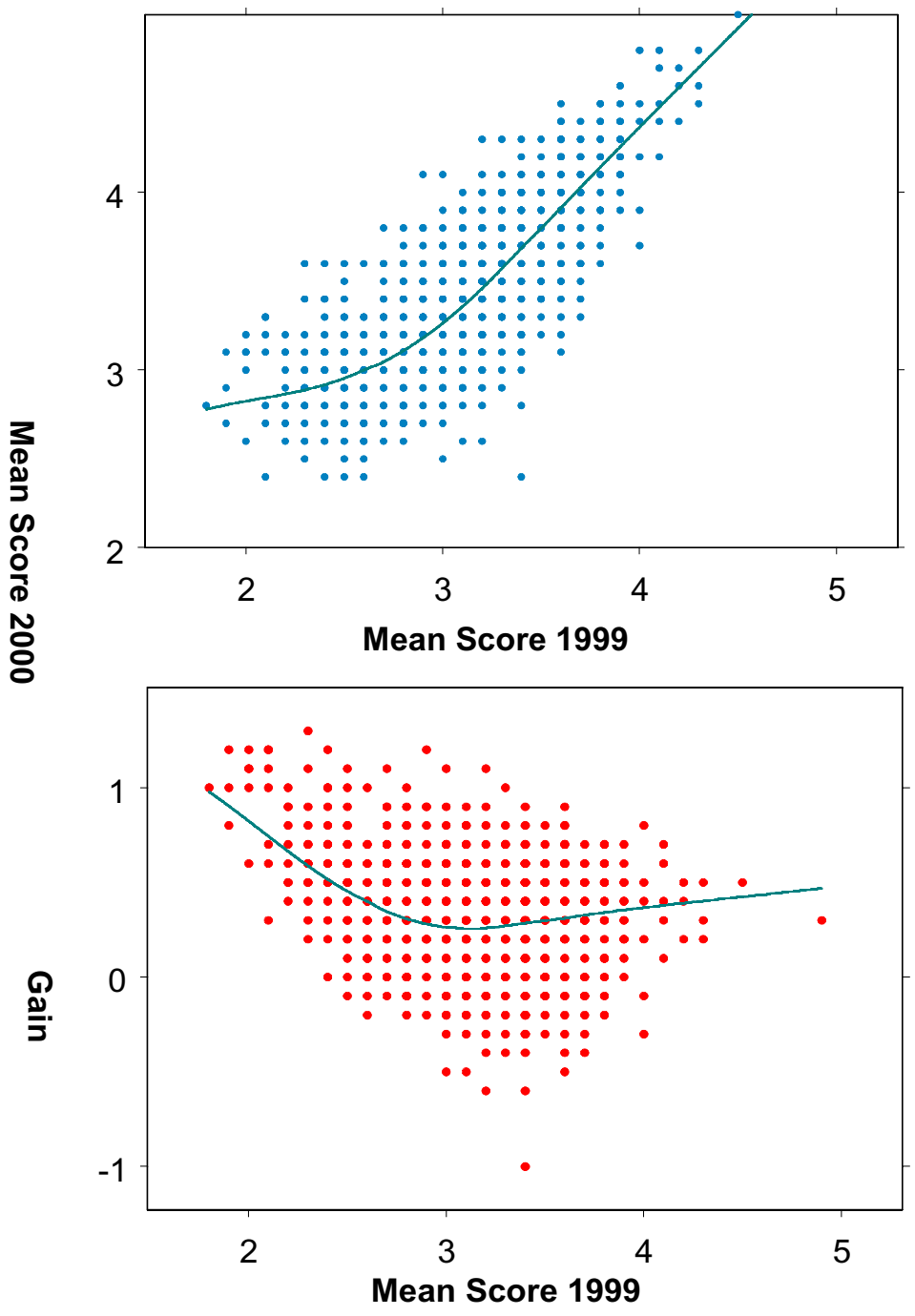

Figure 4. Writing 2000 scores and gains as a function of 1999 scores.

schools, the writing test has helped them avoid an F rating." My findings are consisted with this explanation.

The pattern of score improvements on the FCAT should give Florida officials pause and trigger a serious research effort to identify potentially harmful imbalances and deficiencies in the A-Plus program. Until a far better understanding of and experience with the Florida accountability system is at hand, Greene's generalization from the Florida data he examined to the desirability of a nationwide implementation is premature. It appears that the program's strong attention to the lower portion of the score distribution and the aggressive efforts to improve test scores in that region have produced substantial unintended consequences. Much more evidence is needed to arrive at a sufficiently detailed account of the program's operations and impact. The short list will include documentation of instructional practices in response to the incentive system in place for high- and low-scoring schools, an examination of the implementation and utility of school improvement plans, and data on possible program effects on retention, dropout, and interschool mobility patterns.

Greene's report leaves open the question of the extent to which a "voucher threat" was the key to score improvements in F schools. But even if vouchers were a dominant factor in motivating failing schools to act, the action they produced cannot be considered desirable by anyone who aims to "raise the bar" for students and schools. A minimum performance level in writing should not be considered a worthy educational goal for an ambitious accountability system such as the Florida A-Plus program. Yet, this appears to be the main achievement of the program in F schools. Coupled with a pattern of stagnation in other grade groups, especially in reading, these findings point to aspects of the program that deserve closer scrutiny. However, the reader of the Manhattan Institute report is offered a sense of the program's being a success. It is, therefore, appropriate to recall Cronbach's (1980) advice to the evaluator:

> Disillusion is the bitter aftertaste of saccharine illusion. It is self-defeating to aspire to deliver an evaluative conclusion as precise and as safely beyond dispute as an operational language from the laboratory. . . . When the evaluator aspires only to provide clarification that would not otherwise be available, he has chosen a task he can manage and one that have social benefits. (p. 318)

References

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

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

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

Greene, J. P. (2001a). *An evaluation of the Florida A-Plus Accountability and School Choice Program.* New York: The Manhattan Institute. Available February 27, 2002, from www.manhattan-institute.org / html / cr_aplus.htm

Hegarty, S. (2000, June 21). Why are Florida children writing to much better? *St. Petersburg Times.* Available February 26, 2002, from www.sptimes.com / News / 062100 / State / Why_are_Florida_child.shtml