ARIZONA STATE UNIVERSITY

Colorado
University of Colorado at Boulder

EPRU | **EDUCATION POLICY RESEARCH UNIT**

EPIC Education and the Public Interest Center

DOCUMENT REVIEWED:	Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure
AUTHORS:	Cecilia Elena Rouse, Jane Hannaway, Dan Goldhaber, and David Figlio
PUBLISHER/THINK TANK:	Urban Institute's National Center for Analysis of Longitudinal Data in Education Research
DOCUMENT RELEASE DATE:	November, 2007
REVIEW DATE:	January 15, 2008
REVIEWER:	Damian W. Betebenner
E-MAIL ADDRESS:	dbetebenner@nciea.org
PHONE NUMBER:	(603) 516-7913
EPSL DOCUMENT NUMBER:	EPSL-0801-249-EPRU

Summary of Review

This study examines the relationship between high-stakes school accountability and its effects upon student test scores and school policies. The authors seek to understand the extent to which accountability sanctions and incentives for the poorest-performing schools in Florida explain subsequent changes in school practices and policies as well as achievement — measured by state assessment data, Stanford-10 assessment data and surveys of public school principals. Based on statistical analysis of the lowest-performing schools, the authors report that accountability incentives and sanctions are related to school practice and policy as well as to student achievement. The report uses comprehensive data sources and applies appropriate methodologies to address the research question. Its analyses demonstrate a mediating relationship for school policies between accountability and achievement gains, a finding consistent with both the literature on the subject and common sense. However, the report overstates and makes causal claims about the relationship between accountability sanctions and improvements in school achievement. In this way, the report's title and some causal statements in the body of the report are unfortunate in that they overstate the report's sound findings and suggest that vouchers and other accountability measures are shown to be the cause of achievement gains in some of Florida's lowest-performing schools.

Review

I. INTRODUCTION

Do high-stakes accountability systems yield higher student achievement, particularly for schools facing sanctions? A de facto answer of yes was embedded in No Child Left Behind (NCLB), which mandated adoption of high-stakes assessments as a means for driving education reform. Not until recently, however, have researchers had the rich system-level data with which to examine the impact of large-scale accountability reforms on the achievement of students at the state level.

If higher achievement follows establishment and implementation of sanction- and incentive-laden accountability systems, what is the mechanism behind the increase? Do schools facing sanctions alter and modify their practices and policies to avoid harsher consequences associated with continued low performance? Do these modified practices and policies lead to greater efficiency that manifest in the form of higher student achievement? The premise is plausible.

Numerous researchers have documented that school practices are often altered in perverse ways (e.g., teaching to the test, narrowed curriculum, and cheating) in response to high-stakes accountability.[1] Clearly, the responses to the increased stimulus of accountability pressure do not necessarily conform to best practice. Are such undesirable consequences a *fait accompli*?

The Urban Institute Working Paper reviewed here is titled "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure," and authored by Cecilia Elena Rouse, Jane Hannaway, Dan Goldhaber, and David Figlio.[2] It presents an analysis of the Florida A+ accountability system. The authors challenge the notion that all schools facing accountability sanctions will try to "game the system" by making superficial changes that result in higher test scores but not greater learning. They explore principals' self-reported reactions to very low state ratings and the threat of sanctions, attempting to unpack the "black-box" mediating between the accountability system and student achievement. They find that schools receiving a grade of "F" in the Florida A+ Plan for Education and confronted with the incentives and sanctions associated with such a grade responded by altering practice and that these changes explained, at least in part, subsequent gains in student achievement.

II. FINDINGS AND CONCLUSIONS OF THE REPORT

The report seeks to explain the impact on student achievement and school policy of the receipt of grade of F under Florida's A+ Plan for Education. In particular, the authors examine 35 elementary schools that received an F. Under the plan, these schools faced a combination of incentives and sanctions, including outside evaluation by a community assessment team, technical assistance from the Florida Department of Education, and supplementary assistance through Florida's Assistance Plus program, as well as the possibility for student transfer using vouchers, called "Opportunity Scholarships." The authors take special care to focus on these schools and compare their subsequent performance to other low-performing schools — those that received a D and therefore faced a slightly different combination of incentives and sanctions.

The authors contend that incentives and sanctions might improve school performance by changing the policies of key school administrators. The theory of action here is that intensive scrutiny and supervision under the Florida A+ Plan, personal drive to remediate low performance, and avoidance of the stigma of administering to an "F" school will drive school-level changes in policy and practice, which then leads to higher student achievement. The sanctions and incentives of being labeled an "F" school should alter practice and policy and do so in ways that truly benefit the students and do not simply try to game the system.

The authors pursue two analytic paths to explore these proposed relationships: (1) statistical analyses of the impact of an F on student test scores, and (2) statistical analyses of the impact of an F on school policies. The combined analyses seek to unpack how policy mandates are mediated by school policy changes to affect increases in student achievement.

Data for the analyses come from two sources: (1) administrative data including individual student data with characteristics such as race, sex, ELL and disability status, as well as both high-stakes FCAT test scores and low-stakes Stanford-10 test scores; and (2) survey data of principals of all public schools in Florida, conducted in the 2001-02 and 2003-04 school years.[3] The data sets used for the analyses are unique and impressive given their scope and breadth. More importantly, the data sets are adequate to address the research questions posed in the report.

As described above, the two types of analyses parallel these two sets of data and investigate the relationship between accountability sanctions and subsequent changes in student achievement. In examining the impact of an F on student test scores, the authors use regression techniques on cross-sectional and multi-cohort data, controlling for a variety of student and school characteristics. According to the researchers, the analyses seek to establish the "plausibly causal" (p. 14) impact of the receipt of a grade of F on student test score gains. Technically, and as these authors recognize, it is not the F grade that causes student test score gains but the practices motivated by such a grade.

The report attempts to control for the possibility that the changes reflected in assessment score outcomes do not reflect an improvement in students' broader knowledge base but rather their ability to do well on a specific assessment (i.e., that students were taught to the test). For this purpose, the authors investigate results for both the high-stakes FCAT test as well as the low-stakes Stanford-10 assessment. This was a sensible approach, but given the highly standardized nature of the Stanford-10 assessment, the authors should probably not rule out teaching to the test as a contributing factor solely on the basis of consistent gains across both the FCAT and Stanford-10. Specifically, any skills acquired through teaching to the FCAT such as improved test-taking strategies would likely transfer well to the Stanford-10. Subsequent analyses using the principal survey data indicate that achievement gains are only modestly attributable to narrowing the curriculum to focus on test performance.

The report finds that across FCAT and Stanford-10 reading and mathematics exams, students in F-graded schools demonstrated larger test score increases than did students in other schools. Since students in "F" schools have very low achievement, this finding is not surprising and in fact is consistent with regression to the mean.[4] It is unfortunate that the authors report this in-

formation separately from later analyses where they show that the increases observed are statistically significant even after accounting for these regression effects.

Because of the large number of schools in the dataset, the authors examine first-time "F" schools separately from repeat "F" schools, and they find consistent and significant differences in test score increases between the two types of schools. In particular, the authors report that as schools remain in "F" status, accountability pressure increases and so do test scores. Given the small sample size associated with this comparison, it is impossible to completely rule out the possibility that these increases were the result of unintended policies such as extensive test prep.

To investigate the impact of "F"-school status on school policies, the authors organized schools' policies into various domains, reflecting types of changes that schools would likely make. The researchers found that schools facing the increased pressures associated with F schools tended to adopt block scheduling, increase time for collaborative planning and class prep for teachers, and otherwise reorganize the school scheduling structure. In addition, the authors found that "F" schools tended to focus on low-performing students, increase the time spent on instruction, and increase resources available to teachers. The authors also found some evidence of narrowing of the curriculum, with an increasing focus on the tested subjects.

Overall, the report's findings are not surprising and reflect a classic mediating relationship. Increased pressure applied to low-performing schools within the Florida A+ Plan led to changes in practice and policy at the school level which in turn accounted for a large proportion of the gains in achievement observed for these schools.

III. REVIEW OF THE REPORT'S METHODOLOGIES

What the report investigates, although never labeled as such, is a "mediating relationship," where school sanctions alter school policies, which in turn increase student achievement. The report uses a variety of appropriate statistical models to estimate the impact of receiving an F grade on student achievement and school policies. As discussed in this review, the methodologies generally support the report's goal of investigating the relationships between policies and changes in student and school academic performance. However, given the observational nature of the data, causal attribution is difficult to establish, as the authors recognize in several parts of the text yet seemingly ignore in others.

IV. REVIEW OF THE VALIDITY OF THE FINDINGS AND CONCLUSIONS

The report's primary strength is its use of comprehensive datasets. The authors thoroughly analyze the data to investigate the impact of accountability sanctions on both student achievement and school policies. Many of the findings of the study are, in fact, consistent with what is found in the existing literature. To address concerns that gains in student achievement might be solely a function of schools "gaming" the system, the authors use multiple achievement tests to show that student achievement gains are consistent and meaningful. As mentioned previously, consistent gains across both the FCAT and Stanford-10 do not completely eliminate the possibility of less beneficial policy changes leading to increases in student test scores. However, the extensive survey data does address many of these concerns, since the researchers are able to corre-

late specific policy changes with the observed test-score gains.

The most prominent shortcoming of the report is its tendency to overstate the predictive relationships indicated by their statistical analyses. The authors have only demonstrated that school policies and practices function as a mediating variable between accountability sanctions and student achievement. The data and methods allow for nothing more, and the last paragraph on page 34 strongly suggests this mediating relationship:

> Across the specifications …, the estimated effect of "F" grade receipt decreases with the inclusion of these policy variables, with percentage reductions that range from very modest to very large. The share of the test score gain associated with "F" grade receipt is at least 15 percent with regard to reading and at least 38 percent with regard to mathematics. Moreover, virtually the entire explained portion of the test score gains associated with an "F" grade is apparently due to the five policy domains that we found to have the strongest cross-sectional relationship with student test score gains.

The use of "associational" language here is more appropriate. Yet scattered throughout the report, the authors write as though the accountability pressure *causes* the improvement in student achievement. For example, on page 22 the authors propose to "further identify the causal effect of receipt of an 'F' grade on student test score improvements…," even though the regression discontinuity analyses do not support such causal attribution.[5] Earlier, on page 14, the authors propose to estimate the "plausibly causal impact of receipt of an F grade on student test score" performance. More broadly, the report's title suggests that vouchers, among all the other accountability provisions of the Florida A+ system, led to increased student achievement and that without these threats, the policies and practices at the school would not have changed. However, there is no supporting evidence that this is true. Moreover, even if it is true that the Florida policy of vouchers plus other accountability provisions did lead to the changes in policy and practice, nothing in this new research allows a policy maker to single out either vouchers or other accountability provisions (or a combination) as having such an effect. By the very fact that different states incorporated difference incentives and sanctions into their accountability system, a different set of sanctions or incentives conceivably might lead to the same achievement outcomes. The title of the report suggests that it might be these components (e.g., vouchers) of accountability system that are responsible, yet the report does not provide evidence for that claim.

V. USEFULNESS OF THE REPORT FOR GUIDANCE OF POLICY AND PRACTICE

As the authors straightforwardly acknowledge in the final paragraph of the report, it is not clear that the findings of Florida will transfer to other locales adopting an accountability program similar to the Florida A+ program. The hypothetical question of interest is whether the A+ Plus accountability system, transplanted to another state, would yield similar results. Given the idiosyncrasies of states, there is no way of definitively answering such a hypothetical question. Perhaps a good way to think about the significance of these findings is that other states should investigate similar questions using their own accountability systems.

As researchers continue to unpack the "black-box" of school accountability, they are certain to find particular school practices, already documented in the literature, that produce the gains in achievement everyone desires. The trick is to identify the policy levers that are most likely to give rise to widespread adoption of effective practices. The new Urban Institute report suggests that one or more elements of the Florida accountability systems may offer such a lever.

However, because changes in school policy and practice can occur for many reasons, this research should not be read to show that the accountability system "led to" or "caused" the student achievement increases. Nor does the new report consider whether the accountability levers in Florida are the most effective means of quickly and beneficially transforming the policies and practices of schools in a way that leads to increased student achievement. The results of this study indicate that in Florida's A+ Program and its associated sanctions led to some of these transformations. Whether or not it is the best or the only means of achieving these ends is the more relevant question.

NOTES & REFERENCES

[1] See Koretz, Daniel (2003). "Using Multiple Measures to Address Perverse Incentives and Score Inflation." *Educational Measurement: Issues and Practice* 22, no. 2: 18-26).

Linn, Robert. (2005) "Alignment, High Stakes, and the Inflation of Test Scores," in *Uses and Misuses of Data for Educational and Accountability Improvement* (ed. by Joan L. Herman and Edward H. Haertel) (Malden, Massachusetts: Blackwell Publishing), pp. 99-118.

[2] Rouse, C. E., Hannaway, J., Goldhdhaber, D., & Figlio, D. (2007). *Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure (Working Paper no. 13)*. Washington DC: Urban Institute's National Center for Analysis of Longitudinal Data in Education Research. This Urban Institute report was published as a Working Paper, and the authors welcomed comments. This review will hopefully be useful to the authors in that regard.

[3] The response rate for the survey was an impressive 70%.

[4] Regression to the mean, called regression toward mediocrity by Galton, denotes the fact that post-observations always are, on average, less extreme than pre-observations. In particular, using standard deviation units, low-achieving students on a pre-test will on average be less extreme in standard deviation units on the post-test.

[5] Regression discontinuity designs (those that compare data from participants assigned to either the program or comparison groups based, in this case, on whether the school is labeled D or F) allow for the attribution of cause only under strict assumptions of a continuous relationship between assignment and outcome variables near the treatment cutoff. This condition is violable if individuals can exert control over values on the assignment variable, which is the case with schools receiving a grade of F.

The Think Tank Review Project is made possible by funding from the Great Lakes Center for Education Research and Practice.

SUGGESTED CITATION:
Betebenner, D. (2008). *Review of "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure."* Boulder and Tempe: Education and the Public Interest Center & Education Policy Research Unit. Retrieved [date] from http://epicpolicy.org/thinktank/review-feeling-florida-heat-how-low-performing-schools-respond-voucher-and-accountability-