American Economic Journal: Economic Policy 2013, 5(2): 251–281
http://dx.doi.org/10.1257/pol.5.2.251

Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure[†]

By Cecilia Elena Rouse, Jane Hannaway, Dan Goldhaber, and David Figlio*

While numerous studies have found that school accountability boosts test scores, it is uncertain whether estimated test score gains reflect genuine improvements or merely "gaming" behaviors. This paper brings to bear new evidence from a unique five-year, three-round survey conducted of a census of public elementary schools in Florida that is linked with detailed administrative data on student performance. We show that schools facing accountability pressure changed their instructional practices in meaningful ways, and that these responses can explain a portion of the test score gains associated with the Florida school accountability system. (*JEL* H75, I21, I28)

There exists an emerging consensus that school accountability systems improve students' performance on standardized tests. Numerous studies have found that the incentives introduced by No Child Left Behind have led to substantial gains in at least some subjects (Ballou and Springer 2008; Dee and Jacob 2011; Ladd and Lauen 2009; Reback, Rockoff, and Schwartz 2011; Wong, Cook, and Steiner 2010), and others have found that accountability systems implemented by states and localities have also improved average student test performance (Carnoy

* Rouse: Education Research Section, Firestone Library, Princeton University, Princeton, NJ 08544, and NBER (e-mail: rouse@princeton.edu); Hannaway: American Institutes for Research, 1000 Thomas Jefferson Street, NW, Washington, DC 20007 (e-mail: jhannaway@air.org); Goldhaber: Center for Reinventing Public Education, University of Washington Bothell, Box 358200, Seattle, WA 98195 (e-mail: dgoldhab@u.washington.edu); Figlio (corresponding author): Institute for Policy Research, Northwestern University, 2040 Sheridan Road, Evanston, IL 60208, and NBER (e-mail: figlio@northwestern.edu). The authors are listed in reverse alphabetical order. This research could not have been undertaken without the help of many people. We thank Edward Freeland, Craig Deshenski, Kenneth Mease, Rob Santos, and Fritz Scheuren for their exceptional help with the survey and sample development. We appreciate the assistance of the Florida Department of Education in providing us both with administrative student data as well as with sampling frames for our survey analysis. Jay Pfeiffer, Jeff Sellers and others at the Florida Department of Education provided very helpful advice regarding Florida education policy and the administrative data used in the analysis. We also thank Emily Buchsbaum, Cynthia Casazza, Sarah Cohodes, Joseph Gasper, Scott Mildrum, Radha Iyendar, Ty Wilde, Grace Wong, and Nathan Wozny for expert research assistance and Jonah Rockoff, Jesse Rothstein, Analia Schlosser, Katherine Strunk, Diane Whitmore Schanzenbach, the anonymous referees, and seminar participants at numerous institutions and conferences for extremely useful conversations and suggestions. Finally, we are indebted to the Annie E. Casey, Atlantic Philanthropies, Smith Richardson and Spencer Foundations, the US Department of Education and the National Institutes of Health for financial support, but the views expressed in this paper do not necessarily represent those organizations supporting this research, the Florida Department of Education, or our host institutions. All errors in fact and interpretation are ours.

† To comment on this article in the online discussion forum, or to view additional materials, visit the article page at http://dx.doi.org/10.1257/pol.5.2.251.

and Loeb 2002; Chakrabarti 2007; Chiang 2009; Figlio and Rouse 2006; Hanushek and Raymond 2004; Neal and Schanzenbach 2010; Rockoff and Turner 2010; Rouse et al. 2007; West and Peterson 2006). The magnitudes of these studies' findings are substantive: Lee's (2008) meta-analysis of 14 cross-state studies finds an average effect size of 0.08, comparable in size to some of the most cost-effective education interventions that have been attempted.[1]

These findings have been treated with some skepticism, however, because there is ample evidence that schools respond to accountability pressures in ways that affect measured performance but may not lead to generalized improvements. For example, many quantitative and qualitative studies indicate that schools respond to accountability systems differentially, allocating resources to the subjects and students most central to their accountability ratings. These authors (e.g., Booher-Jennings 2005; Hamilton et al. 2007; Haney 2000; Krieg 2008; Neal and Schanzenbach 2010; Ozek 2010; Reback 2008; Reback, Rockoff, and Schwartz 2011; Stecher 2002; White and Rosenbaum 2008) indicate that schools subject to accountability pressure focus their attention more on high-stakes subjects, teach skills that are valuable for the high-stakes test but less so for other assessments, and concentrate their attention on students most likely to help them satisfy the accountability requirements.[2] Other studies suggest that schools subject to accountability pressure attempt to shape the testing pool either through exclusions (Cullen and Reback 2006; Figlio and Getzler 2006) or through selective discipline (Figlio 2006). Schools may attempt to artificially boost standardized test scores (Figlio and Winicki 2005) or even manipulate test scores through outright cheating (Jacob and Levitt 2003). The evidence that test score gains associated with accountability systems might not reflect genuine school improvement is therefore ample. These types of behaviors may be the reason that the recent National Research Council panel on school accountability (Hout and Elliott 2011) express a skeptical view about the overall impacts of accountability systems while recognizing the positive gains associated with them.

It is therefore important to gauge the degree to which school accountability is associated with other changes in school policies and practices that could be considered substantive educational responses rather than mere manipulation. To date, however, there exist few studies that attempt to get inside the "black box" of what is going on inside schools. In addition to the literature mentioned above, Goldhaber and Hannaway (2004) present case study evidence from four schools and Chiang (2009) describes conclusions from interviews with five school principals, but the published literature presents no large-scale systematic evidence regarding potential substantive

[1] The 0.08 result refers to the average effect size relative to states that don't have high-stakes accountability systems, and is measured in standard deviations of student performance, where the distribution of student performance is measured across all students in all states. See also Figlio and Loeb (2010) for an in-depth discussion of the school accountability literature.

[2] Stecher (2002) and Hout and Elliott (2011) present evidence of a narrowing of the curriculum to focus on high-stakes subjects at the expense of untested subjects, and Ringwalt et al. (2011) and others describe ways in which accountability-pressured schools have cut programs focused on socio-emotional and health outcomes. Other evidence regarding health outcomes comes from the work by Anderson, Butcher, and Schanzenbach (2011) and Yin (2011) that show school accountability may lead to increased obesity rates, perhaps because of a decreased emphasis on physical education during the school day. However, this does not necessarily mean that student performance in low-stakes subjects suffers in the wake of accountability systems. Greene, Trivitt, and Winters (2009), for instance, find that the gains in reading and math in Florida spilled over to higher science test scores as well.

responses to accountability pressures.[3] Reback, Rockoff, and Schwartz (2011) find that school accountability induces teachers to work harder (and also to narrow the curriculum), but they cannot study other, more difficult-to-measure outcomes, such as specific instructional policies and practices or major changes in the culture of a school.[4] Bacolod, DiNardo, and Jacobson (2009), Chiang (2009), and Craig, Imberman, and Perdue (2013) present evidence that crude expenditure categories are influenced by school accountability pressures, but do not have information regarding what is actually being purchased with these changed resources. In sum, very little is known about what schools are doing in terms of policies and practices that might be leading to the improvements in test performance observed in the data.

The primary reason for the lack of research on this topic is lack of data. While information on student test scores, broad resource allocation categories, and a few crude indicators of student body characteristics (e.g., race/ethnicity, eligibility for the National School Lunch Program, and disability rates) are prevalent, there exist no large-scale datasets that systematically describe school instructional policies and practices. In this paper, we address this gap in the research on school responses to the pressures introduced by school accountability and school choice systems. We do so using a unique and comprehensive dataset we have generated. During the 1999–2000, 2001–2002, and 2003–2004 school years, we administered surveys to every principal of every regular public school in the state of Florida.[5] Our surveys, which were developed based on input from an expert panel of economists, sociologists, political scientists, and education scholars, were designed to help us identify key aspects of schools' instructional policies and practices. And we achieved a consistently high—over 70 percent—response rate on these surveys. In this paper, we focus only on the elementary schools that we surveyed; we do this because the policy and practice variables appear to be the most comparable in this population of schools, and because many of our questions apply primarily to the elementary school environment.

We are able to identify the causal effects of accountability pressures by exploiting an important change to Florida's accountability system that took place in 2002. While the broad features of the change in the grading system were known in advance, the specific parameters of the grading system change implemented by the Florida Department of Education (FDOE)—including how much weight would be placed on proficiency rates versus student learning gains in reading and math, and indeed, how student learning gains would be calculated and used for school accountability—were only determined well into the 2001–2002 academic year. We argue that it was impossible to know exactly where a school would fall along the accountability grading distribution until the actual grading formula was announced. The element of surprise is important because if schools could anticipate the new grading system

[3] See Hannaway and Hamilton (2007) for a review of the case study literature on classroom responses to accountability pressures.

[4] The education literature offers little guidance as to what would be the ideal allocation of resources in a school. There are very few policies and practices for which there exists strong evidence of effectiveness, and the literature developing causal tests of educational interventions is still in its infancy. We therefore are agnostic about the types of responses that would be ideal, and instead are more interested in simply knowing whether we observe any substantive responses at all.

[5] We make use primarily of the two survey rounds most adjacent to the policy change in question.

in advance, they could then make adjustments in an attempt to ensure that they were on the positive side of the grading distribution. This accountability/grading "shock" led some schools to receive higher school grades than they would have received under Florida's previous school grading system, and others to receive lower grades than they would have received under the previous system.[6] The grading shock is well suited for a regression-discontinuity design because the state assigned a set of accountability "points" to each school based on grading criteria, the details of which were unknown at the time that our surveys were administered, but the accountability points were collapsed into five letter grades such that virtually identical schools received very different treatments under the new accountability regime. In the regression-discontinuity models, we control for a function of the number of accountability points that a school received when estimating the effect of receiving a grade of "F," and in addition we control for the grade that the school would have received had the state not changed its grading system. Therefore, we estimate the effect of being just barely below the threshold for receipt of a "F" to being just barely above this threshold, for schools that would have received the identical treatment under the old school accountability regime in Florida. Using this strategy we can identify the instructional responses of schools to changes in accountability pressure associated with idiosyncratic features of the accountability system's rules.

We find evidence that schools change their instructional policies and practices when accountability pressure changes. Similar to other studies, we find that schools faced with accountability pressure appear to focus on low-performing students. But we find other instructional policy and practice results as well: schools appear to lengthen the amount of time devoted to instruction, adopt different ways of organizing the day and learning environment of the students and teachers, increase resources available to teachers, and decrease principal control. We find that the instructional policies and practices most associated with "F" grade status are also associated with the differential test score gains of "F" schools. We do not assert that we have identified a formula for success in low-performing schools; indeed, while we can present causal evidence suggesting that accountability pressure increases test scores and that this same pressure induces schools to change their instructional policies and practices, we do not know which specific policies and practices are responsible for the test score improvements. But the fact that we find these substantive responses to accountability pressure provides strong evidence indicating that at least some of the test score improvements identified by the literature reflect genuine changes in the ways in which schools do business.

[6] Florida began grading schools on an "A" through "F" scale in 1999. The grading system in 1999 was based on the percentages of students meeting proficiency standards in reading, writing and mathematics. The 2002 change in the system introduced not only these proficiency rates but also measures of individual student learning gains from year to year in reading and mathematics, as well as improvements in reading for the school's lowest achieving students. Forty-two percent of Florida schools experienced an "upward shock" in 2002, while nine percent of schools experienced a "downward" shock. The specifics are discussed in greater detail in the next section of the paper. Other researchers have also exploited this grading shock to estimate the effects of school accountability pressures on test scores (Chiang 2009; Rouse et al. 2007; West and Peterson 2006), teacher labor markets (Feng, Figlio, and Sass 2010), and private donations to public schools (Figlio and Kenny 2009).

I. The Florida School Accountability Program

Florida's 1999 A+ Plan for Education introduced a system of school accountability with a series of rewards and sanctions for high-performing and low-performing schools. The A+ Plan called for annual curriculum-based testing of all students in grades three through ten, and annual grading of all public and charter schools based on aggregate test performance. As noted above, the Florida accountability system assigns letter grades ("A," "B," etc.) to each school based on students' achievement (measured in several ways). High-performing and improving schools receive rewards, while low-performing schools receive additional assistance as well as sanctions.

The assistance provided to low-performing schools primarily consists of three components. First, each school district with a school receiving a "D" or "F" is evaluated by a "community assessment team" comprised of local and state leaders, including parents and business representatives. This team makes recommendations to the state and the local school board on how to improve the district's schools. In addition, the Florida Department of Education (FDOE) makes available technical assistance—with needs assessments and the implementation of school improvement plans—to *any* school in the state, with priority given to "D" and "F" schools (as well as those in rural or sparsely populated areas).[7]

The most direct form of assistance available for low-performing schools is through Florida's *Assistance Plus* program. While this program offers no direct funding, it mandates that districts allocate certain resources and targeted funding for "F" schools. In particular, this program requires that "F" schools are also given priority for the *Just Read, Florida!* Program, which provides reading coaches trained in scientifically based reading research to the lowest performing schools (the program was funded at $11 million in 2001–2002). It also includes several other forms of assistance that increased school oversight: a team of specialists (Assistance Plus Teams), assigned to each "F" school, were designated to help structure improvement efforts; state-coordinated regular conference calls and reports provided continuous progress-reporting and feedback to schools and districts; curriculum assessment and course materials were aligned; data specialists analyzed student achievement data to provide technical assistant for schools and instructors; and administrators made recommendations regarding professional development for teachers.[8] Importantly, the state does not dictate a formula for school improvement, but rather encourages a locally developed response. Each year the state identifies a set of successful schools serving a range of different student background characteristics and maintains a list of policies and practices that leaders of these schools argue work in their circumstances. This list is very comprehensive, and successful school testimonials encompass the set of policies and practices that we include in our surveys.

While "F" schools receive priority through Florida's *Assistance Plus* program, "D" schools were also targeted. In fact, in the 2007–2008 academic year (the only

[7] The 2002 Florida Statute, Title XLVIII ("K–20 Education Code"), Chapter 1008.345 (Assessment and Accountability; Implementation of state system of school improvement and education accountability). Downloaded from http://www.leg.state.fl.us/statutes on November 16, 2007.

[8] These statements are based on documents received from the FDOE regarding the program in 2001–2002 (they can also be accessed at http://schoolgrades.fldoe.org/pdf/0102/assist_plus.pdf).

year for which we were able to obtain this information), 79 "F" schools and 220 "D" schools received such assistance, the approximate number of potentially eligible "F" and "D" schools. As such, we suspect that while "F" schools receive priority, "D" schools heavily benefit from this state aid as well.[9]

On the sanction side, the most controversial and widely publicized provision of the A+ Plan was the institution of vouchers, called "Opportunity Scholarships," for students attending (or slated to attend) chronically failing schools—those receiving a grade of "F" in two years out of four, including the most recent year. These "Opportunity Scholarships" allowed students to attend a different (higher rated) public school, or an eligible private school.[10] Poor-performing schools were also subject to additional state and district scrutiny and oversight. All "D"- and "F"-graded schools are subject to site visits and required to send regular progress reports to the state, such as through the requirements of the *Assistance Plus* program.[11]

School grading began in May 1999, immediately following passage into law of the A+ Plan. Between 1999 and summer 2001, schools were assessed primarily on the basis of aggregate test score *levels* (and also some additional nontest factors, such as attendance and suspension rates, for the higher grade levels) and only in the grades with existing statewide curriculum-based assessments,[12] rather than on the *progress* schools made toward higher levels of student achievement.

Starting in summer 2002 school grades began to incorporate test score data from all grades from three through ten and to evaluate schools not just on the level of student test performance but also on the year-to-year progress of individual students. However, while at the beginning of the 2001–2002 school year several aspects of the new school grading formula that was to be used to assign grades in summer 2002 were known (i.e., school grades were to be based on test scores from all students in all tested grades; the standards for proficiency in reading and mathematics were to be raised; and school grades would incorporate some notion of student learning gains into the formula), the specifics of the formula that would put these components together to form the school grades were unknown until the spring of 2002.[13] A key component of our analytic strategy is to take advantage of the fact that during the 2001–2002 school year, schools could not necessarily anticipate their school

[9] We note, as well, that in documents referenced above the FDOE highlights assistance given to "D" as well as "F" schools, reinforcing our belief that "D" schools heavily benefit from such assistance.

[10] Private schools participating in the program were required to accept the Opportunity Scholarships as the full cost of tuition. Religious schools were permitted to participate in the Opportunity Scholarship Program so long as they did not require scholarship recipients to participate in religious programs or religious education activities.

[11] Details on oversight and reporting requirements can be found online at http://www.bsi.fsu.edu/PerformanceUpdates/performanceupdates.aspx.

[12] Students were tested in grade 4 in reading and writing; grade 5 in mathematics; grade 8 in reading, writing, and math; and grade 10 in reading, writing, and math.

[13] As an example, it was not determined until late in the school year that in the new system, students who were already deemed proficient were considered to be automatically making "learning gains" for the purposes of school grading, regardless of whether their actual scores were advancing or falling backward. The treatment of the measured learning gains of proficient students makes a tremendous difference in the actual grades awarded. We calculate that between 12 and 27 percent of the schools would have received a different school grade in 2002 depending on which reasonable alternative treatment of learning gains for proficient students would have been applied. And this is only one of the numerous specifics that the state had to work out in 2002; there were other formula elements, such as the differential emphasis on low-performing readers that were being discussed throughout much of the school year. Clearly it would have been very difficult for schools to anticipate and react to these changes as they were still being debated.

Table 1—The Distribution of School Grades, by Year

School grade	School year					
	Summer 1999	Summer 2000	Summer 2001	Summer 2002	Summer 2003	Summer 2004
	All schools					
A	183	552	570	887	1,235	1,203
B	299	255	399	549	565	515
C	1,180	1,115	1,074	723	533	568
D	565	363	287	180	135	170
F	70	4	0	60	31	34
N	0	0	76	102	2	0
Total	2,297	2,289	2,330	2,501	2,501	2,490
	Elementary schools					
A	119	485	389	623	928	974
B	214	180	324	368	360	333
C	713	614	636	452	299	284
D	448	260	215	124	63	67
F	61	4	0	35	18	9
N	0	0	46	68	2	0
Total	1,555	1,543	1,610	1,670	1,670	1,667

Source: Authors' calculations from state data.

grade in summer 2002 because the specific changes in the grading formula were not decided until well into the school year. Thus, many received what we believe to be an accountability "shock," enabling us to make a pre-post comparison of changing degrees of accountability pressure. We exploit the grade change both by estimating a regression discontinuity model in the grade points model used to determine school grading, as well as by controlling for the grade that the school would have received had the accountability system not changed. This latter feature helps to ensure that our results are not being driven by unobserved differences in the attributes of schools of different grade levels. That said, to the degree to which policy changes were anticipated, any anticipatory reactions would work to bias our results toward a finding of zero reaction to the grade change.

Table 1 shows the distribution of schools across the five performance grades for the first six rounds of school grading, for all graded schools in Florida, both for the entire population of schools and for elementary schools—the focus of our analysis. As is apparent from the variation across years in the number of schools that fall into each performance category, there are considerable grade changes that have taken place since the accountability system was adopted. Most notable is the fact that while 70 schools received an "F" grade in the first year (1998–1999), only four did so the subsequent year and none did in summer 2001. At the same time, an increasing number of schools were receiving "A"s and "B"s. This is partly due to the fact that schools had learned their way around the system: a school had to fail in all three subjects to earn an "F" grade, so as long as students did well enough in at least one subject the school would escape the worst stigma. Goldhaber and Hannaway (2004) and Chakrabarti (2007) find evidence that students in failing schools made the biggest gains in writing, which is viewed to be one of the easier subjects in which to show

Table 2—Transition Matrix in Predicted Grades Based on 2002 Grade Change, Elementary Schools (*row percentages*)

		Grade in 2002 based on new (summer 2002) grading system				
		A	B	C	D	F
Simulated grade in 2002 based on old (summer 2001) grading system	A	0.89 [264]	0.11 [33]	0.00 [0]	0.00 [0]	0.00 [0]
	B	0.48 [188]	0.38 [149]	0.13 [50]	0.00 [1]	0.00 [0]
	C	0.23 [167]	0.25 [181]	0.46 [331]	0.05 [37]	0.00 [2]
	D	0.02 [3]	0.01 [2]	0.38 [69]	0.44 [79]	0.15 [28]
	F	0.00 [0]	0.00 [0]	0.00 [0]	0.29 [2]	0.71 [5]
Fraction upward shocked		0.58	0.50	0.15	0.02	0.00
Fraction downward shocked		0.00	0.09	0.11	0.32	0.86
Fraction unshocked		0.42	0.41	0.74	0.66	0.14

Notes: All row percentages are student-weighted with the number of schools in brackets. Simulated grade changes are generated by applying both the old grading system and the new grading system to 2002 student test scores. They are therefore generated based on precisely the same student tests; the only differences in the calculations are the formulas used to convert these same tests into school grades.

improvement quickly. When the rules of the game changed, so did the number of schools caught by surprise. For example, the number of schools earning an "F"grade increased to 60 in summer 2002, and the number of "A" schools grew as well, largely because the "grade points" system of school grading allowed schools that miss performance goals in one area to compensate with higher performance in another.[14]

A way to judge the extent to which schools might have been caught "off guard" by the new system is to compare the grades that schools actually received in summer 2002 with the grade that they would have been predicted to receive in 2002 based on the "old" grading system (that in place in 2001). In Table 2, we make this comparison for all graded elementary schools using the full set of student administrative test scores provided to us by the FDOE.[15] It is possible to make these calculations because the summer 2001 grading system's formula is known. We estimate that about 52 percent of elementary schools experienced no change in their school grade based on the change in the accountability system. The other 48 percent either received a higher or lower grade than they might have "expected." For example, two of the seven schools that might have expected to receive an "F" under the old regime received a "D" under the new one (none of these schools received a higher grade). Importantly, 15 percent of schools that might have expected to receive a "D" under the old system received an "F" under the new one. We interpret the difference between these "simulated" grades in 2002 based on the former system and the actual grades received in 2002 as a measure of the change in the grading system per se,

[14] Note that as schools have adapted to the new grading system, the number of "F" schools has also decreased.

[15] The number of observations in Table 2 does not exactly match that in Table 1, because the administrative data on students provided by the FDOE used to simulate each school's grade in 2002 does not include some students in charter schools and "alternative" schools.

rather than the result of behavioral changes on the part of schools or demographic shifts in the school population.

While the accountability system provided incentives for all schools to improve student performance, the sharpest incentives for improvement were faced by the lowest-rated schools. Schools rated "F," for instance, were called out by the FDOE and the news media, as well as school districts. "A"-rated schools received praise as well as financial bonuses, and those in the middle of the distribution received much less notice. We therefore focus attention on the roughly 2 percent of schools that received an "F" (and therefore became voucher threatened, and faced new stigma and state oversight). Among these 35 elementary schools, 6 became voucher eligible.[16,17] Table 3 shows that in 2002 these "F" (elementary) schools were slightly smaller than the higher-rated schools, had a higher proportion of black students and a higher proportion of students eligible for a free- or reduced-price lunch. However, they also spent slightly more per regular-education student and per special-education student compared to schools that received an "A," "B," or "C" grade. The "F" schools look much more similar, demographically and financially, to those schools that received a "D" grade.[18] Given that "D" and "F" schools serve such similar populations, and since the existing research on test score and other responses to the 2002 Florida grading system indicate that the "D"–"F" contrast is the most meaningful, these two sets of schools are the most appropriate to compare.

II. Why Might Incentives Change School Practice?

School administrators have incentive to increase student achievement for a number of reasons. For example, evidence suggests that they are rewarded for improved performance through increased odds of getting a higher-paying principalship or becoming a superintendent (Cullen and Mazzeo 2007). Moreover, if large numbers of students leave a school, the principals and teachers may be left with the less-motivated (harder-to-teach) students, or the school may become so small that it closes and they lose their jobs. This threat, however, is not necessarily powerful since it is also possible that if dissatisfied students leave, management of a school becomes easier. And, if take-up of choice options is low, then the risk of job loss due to lack of demand for services is likely to be low as well.[19]

[16] Among all public schools, 60 received an "F" in summer 2002 and eight of them became voucher eligible. Students in general did not take up the vouchers: Only about 5 percent of voucher-eligible students left their public schools for a private school voucher. A similar fraction left their "failing" schools for higher-rated public schools in their district.

[17] Note that in Table 1 there are 68 elementary schools that received a grade of "N" in 2002. A grade of "N" was supposed to have been given only to new schools, and the state was to assign grade points, but not a formal grade. (We note, however, that 7 of the 68 schools awarded a grade of "N" were not new schools and were perhaps awarded a grade of "N" in a manner contrary to the state's official policy.) We tried three ways of handling these schools: excluding them from the analysis; controlling separately for them; and imputing the school grades that would have occurred had there been an official letter grade (which is the approach we show in the paper). The results are of similar magnitude and statistical significance levels under all three approaches.

[18] This comparison provides additional support for comparing "F"-graded schools to those receiving a "D" grade which we present in Table 6.

[19] Goldhaber (2009) suggests that incumbent teachers and administrators are often not threatened with job loss when schools lose students because the loss of students, and consequently funding, can be managed through normal teacher attrition. Unfortunately, it is difficult to measure the precise take-up rate of the Opportunity Scholarship

Table 3—Average School Characteristics by the School's Accountability Grade in 2002

Variable	School grade in 2002		
	A/B/C	D	F
Number of students	728	624	515
	[258]	[220]	[269]
Proportion black	0.22	0.58	0.74
	[0.22]	[0.29]	[0.27]
Proportion Asian	0.02	0.01	0.004
	[0.02]	[0.02]	[0.01]
Proportion Hispanic	0.18	0.18	0.12
	[0.22]	[0.21]	[0.19]
Proportion Native American	0.003	0.002	0.001
	[0.005]	[0.007]	[0.002]
Proportion mixed race	0.02	0.01	0.01
	[0.02]	[0.01]	[0.01]
Proportion free or reduced-price lunch eligible	0.51	0.81	0.79
	[0.24]	[0.15]	[0.19]
Proportion gifted	0.04	0.02	0.01
	[0.05]	[0.02]	[0.02]
Proportion limited English proficiency	0.09	0.11	0.09
	[0.11]	[0.11]	[0.12]
Proportion classified disabled	0.16	0.15	0.17
	[0.06]	[0.06]	[0.07]
Stability rate	0.93	0.91	0.90
	[0.03]	[0.03]	[0.04]
Special education expenditures per pupil	$8,436	$9,654	$9,604
	[2,098]	[2,178]	[2,097]
Regular education expenditures per pupil	$4,405	$5,327	$5,720
	[845]	[1,162]	[1,113]
At-risk students expenditures per pupil	$5,509	$5,695	$4,929
	[3,774]	[3,751]	[4,233]
Vocational education expenditures per pupil	$102	$217	$465
	[829]	[1,300]	[2,714]

Notes: Standard deviations within grade-specific groupings in brackets. All means are based on data from the 2001–2002 school year. The means for the racial composition of the schools represent the average across grades among test takers; they were calculated from administrative data on individual students. The other characteristics are based on administrative data from the FDOE.

There are several particular elements of the A+ Plan that may motivate school administrators to work hard to improve student performance in Florida. Principals in "F" schools are subject to intensive scrutiny and supervision.[20] For example, as discussed earlier, under the *Assistance Plus* program principals are required to file regular

program because the set of students who are eligible differs from year to year and depends in some cases on prior take-up. According to personal correspondence with FDOE personnel, Opportunity Scholarship take-up rates among newly eligible potential participants ranged from five to ten percent, depending on the year, over the course of the program's operations.

[20] An increase in managerial control is likely to occur in many other settings as a result of poor performance as well. Simon (1957) argued that workers in any organization attempt to restrict the "zone of acceptance" that surrounds their position. The zone establishes the boundaries of managerial control that they expect over their behavior as part of their employment contract. Poor performance, however, can increase the boundaries of the zone inviting greater management control and reducing discretion for the employee.

improvement plans and status reports with the state. The long and detailed reports on the "corrective actions" that the school and district are taking to improve the school suggest that, at a minimum, being labeled as a "low-performing school" brings with it many more bureaucratic "headaches" and increased oversight. Principals and teachers are able to hold off such managerial scrutiny if they increase student achievement.

In addition, as discussed above, school performance might be linked to job security. There is little direct evidence of this,[21] but the threat of job loss may itself be a compelling reason for principals to attempt to institute policies that would improve student achievement. Third, even if school administrators do not fear the increased oversight for the reasons discussed above, they may experience a loss of utility in at least two other ways that would induce them to make an effort to increase student achievement. First, principals and teachers have chosen a profession where the ideal behavior is to foster the development of children. Presumably conforming to this behavior is part of who they are; it is part of their identity. Evidence suggesting that a principal or teacher is not living up to the ideal would contribute to a loss in what Akerlof and Kranton (2005, 2007) refer to as "identity utility." Second, the public nature of the school grades could lead to a loss of status in the wider community. Indeed, based on interviews and focus groups with principals and teachers, Goldhaber and Hannaway (2004) suggest that principals and teachers considered a school grade of "F" as a social stigma. Certainly, principals and teachers in schools with "D" grades might face similar incentives to avoid receipt of an "F" grade in the future, but the desire to avoid an "F" grade is arguably stronger for those who have experienced the grading stigma and increased reporting requirements firsthand.

III. Empirical Strategy and Data

A. *Empirical Framework*

Our empirical framework aims to estimate the effect of having received an "F" grade on school policies. Florida's school grading system change in 2002 lends itself well to a regression-discontinuity design; schools are graded based on their number of "grade points" and grade cutoffs are strictly enforced. In the 2002 grading system that we exploit for identification, a school grade is the sum of (i) the percentage of students meeting proficiency targets in reading, (ii) the percentage of students meeting proficiency targets in math, (iii) the percentage of students meeting proficiency targets in writing, (iv) the percentage of students making gains in reading, (v) the percentage of students making gains in math, and (vi) the percentage of the bottom quarter of the school's student body making gains in reading. That is, a school may earn as many as 600 grade points in the 2002 system. There is therefore a strict discontinuity at, say, the 280 grade point threshold where schools with 279 points receive "F" grades and those with 280 points receive "D" grades. We concentrate on the bottom of the school grading distribution because our school

[21] In Chiang (2008), the author reports there is no difference in principal turnover between "D" and "F" schools. In Table 11A, he shows that 15.0 percent of "F" schools experienced principal turnover, as compared with a statistically indistinguishable 14.7 percent for "D" schools.

surveys were designed primarily to measure the types of school responses that struggling schools might pursue; it turns out ex post that the largest apparent school responses, as measured by test score improvements, were at the bottom of the grading distribution as well. In our analysis, we drop schools that did not receive any accountability points in summer 2002 (364 schools), one school whose grade was set at "NA," and seven schools that were not part of our 2002 sampling frame; this leaves a maximum sample size of 1,659 elementary schools.[22,23]

We estimate school-level models of the form

$$(1) \qquad P_{s,04} = a + bF_{s,02} + c(g(POINTS_{s,02})) + dX_{s,02} + (e_{s,04}),$$

where $P_{s,04}$ indicates whether school s implemented policy P in the 2003–2004 academic year; $F_{s,02}$ is a dummy variable indicating whether or not the school received a failing grade of "F" in summer 2002; $g(POINTS_{s,02})$ is a cubic in the number of grade points the school earned in summer 2002;[24] $X_{s,02}$ is a vector of school-level variables, including dummy variables indicating the school's 2002 "simulated grade," the 2002 value of the policy, the school's estimated expenditures per student in 2002 for special education, vocational education, education for "at-risk" students, and "regular" education, the number of students in the school, the percentage of the school's student body that is classified as disabled, eligible for free- or reduced-price lunch, limited-English proficient, and gifted, and the "stability rate" of the school;[25] $e_{s,04}$ is a normally distributed error term.[26]

The key parameter of interest is b − a school's policy response to having received an "F" grade.[27] Thus, we estimate whether the policy responses of the "F"-graded schools are significantly different from those of higher-graded schools. In some specifications, we also compare the "F" schools to the "D" schools. Note that this comparison allows us to better net out those policy choices dictated by the state-mandated assistance, as "D" schools also heavily benefited from such interventions.

The survey we developed and use for this study (discussed below) contained many questions aimed at understanding the policies and practices that schools

[22] Schools that do not have reported accountability points are exempt from the accountability system either because they have too few students for the state to report test scores, even in aggregate form, or because they are a first-year school. The one school with the grade of "NA" had 467 accountability points (which would have earned that school an "A") however the grade, which was originally set in 2002, has been withdrawn by the State.

[23] The results are not sensitive to restricting the analysis to only those schools earning grades "D" or "F." We report the results of both sets of models to assess the degree to which the support used for the regression discontinuity model affects the outcomes. Both sets of models estimate the effect of receiving an "F" versus a "D" grade.

[24] We note that the results are robust to using a linear, or quadratic functional form for the running variable. When we estimate a more heavily saturated model (a quartic functional form) a few of the point estimates become statistically insignificant although the general pattern of results remains. These results are not surprising given there are only 35 F-graded schools.

[25] The stability rate of a school is the fraction of students who were present in the fall census of students who were still at the same school in the spring census.

[26] We have also estimated models with a richer set of grade cutoffs, but focus on the "F" cutoff for ease of explanation.

[27] We have also estimated the effects at different grading discontinuities. As with previous authors exploiting the 2002 grading shock (Feng, Figlio, and Sass 2010; Figlio and Kenny 2009; West and Peterson 2006) who find that responses to the accountability system are concentrated nearly exclusively at the bottom of the grading distribution, we find little evidence that the other discontinuities are particularly consequential in terms of school policy and practice changes. We therefore concentrate our attention at the "D"–"F" discontinuity.

utilize. Because it would be cumbersome and difficult to digest the pattern of coefficient estimates that would be derived from so many regressions, we attempt to summarize the results by grouping questions into "domains" that are designed to capture related policies.[28] Grouping policies in this fashion also reduces the severity of what we refer to as the "budget constraint problem." Specifically, we assume that school superintendents and principals consider which policies to enact subject to a budget constraint. As such, one would not expect that schools adopt *all* possible policies; rather they pick and choose from some feasible set. For example, a principal may attempt to increase instructional time by sponsoring summer school, an extended school year, or after-school tutoring, but not all three. Grouping similarly intended policies may reduce this problem by allowing us to identify if the administrators adopt a type of policy rather than a specific one.[29]

To see how we analyze the effect of an "F" grade on a "domain," note that we can first rewrite equation (1) to obtain an effect of the accountability system for each policy, where k refers to the kth policy:

$$(2) \quad P_k = a_k + b_k F_{t-2} + c_k(g(POINTS_{t-2})) + d_k X_{kt-2} + (e_{kt}) = W\Theta_k + \nu_k.$$

We combine the estimates using a seemingly unrelated regression (SUR) approach (Kling and Liebman 2004). This approach is similar to simply averaging the estimated effect of receiving an "F" grade on school policies if there are no missing values and no covariates.[30]

More specifically, we first estimate equation (2) (or variants) and obtain an item-by-item estimate of b (i.e., b_k). We then standardize the estimates of b_k by the standard deviation of the outcome using the responses from the 2002 survey for all (elementary) schools (σ_k). Our estimate of the effect of the school accountability system on school policies is then the average of the standardized b's within each domain,

$$(3) \qquad b_{AVG} = \frac{1}{K}\sum_{k=1}^{K}\frac{b_k}{\sigma_k}.$$

To obtain standard errors for b_{AVG}, we need to account for the covariance between the estimates of b_k within each domain. To do so, we estimate the following seemingly unrelated regression system

$$(4) \qquad P = (I_K \otimes W)\Theta + \nu \qquad P = (P_1', \ldots, P_k')',$$

[28] We constructed this grouping of domains from our survey which was designed with the assistance of a technical advisory panel consisting of leading experts in education policy, economics, political science and sociology. As an alternative mechanism of grouping policies we conducted a factor analysis to identify empirically determined principal components. The results were surprisingly similar. That said, we prefer our approach because it is possible to identify with clarity the policies that contribute to each domain.

[29] In addition, we also tried capturing a school's adoption of policies within each domain as the sum of the number of "sub-policies" they adopt. These results are qualitatively similar but less precise than those presented below.

[30] We recognize that the multiple comparison problem is not necessarily solved by this approach because there are still multiple policy categories for which effects are being estimated.

where I_K is a K by K identity matrix and W is defined as in equation (2). We calculate the standard error of the resulting summary measure as the square root of the weighted sum of the variances and covariances among the individual effect estimates. One potential advantage of the SUR is that while estimates of each b_k may be statistically insignificant, the estimate of b_{AVG} may be statistically significant due to covariation among the outcomes.

We present estimates of the summary measure (i.e., the outcomes grouped together within a domain) as well as some of the original underlying regressions.[31] In addition, we also present results using a simple average of the individual items within each domain (an "index" or "standardized mean value"), where the original data have been normalized by the mean and standard deviation of the variable in 2002.[32]

B. *Data*

We carry out our analysis using administrative data that is linked to an original survey of principals of all "regular" public schools in Florida.[33] The administrative data are derived from the Florida School Indicator Reports, which are published online at http://www.fldoe.org/eias/eiaspubs/fsir.asp.

School surveys were conducted three times: in 1999–2000, 2001–2002, and 2003–2004. In these surveys, we asked principals to identify a variety of policies and resource-use areas which, as mentioned above, we have divided into several domains: policies to improve low-performing students, lengthening instructional time, reduced class size for subject, narrowing of the curriculum, scheduling systems, policies to improve low-performing teachers, teacher resources, teacher incentives, teacher autonomy, district control, principal control, and school climate.

We focus our analysis on differences between the 2001–2002 and 2003–2004 survey rounds; we do not present results including the 1999–2000 survey year because the ordering of the questions was different in the 1999–2000 survey versus those in the later two survey rounds.[34] We achieved response rates that exceeded 70 percent in each year. For the analysis, we focus on elementary schools that received a grade in 2002 and were part of our sampling frame. Of the 1,659 eligible schools in the 2004 sampling frame, we achieved an overall response rate of 75 percent. The response rate was 76 percent for schools that received an "A," "B," or "C," in 2002 and 69 percent and 66 percent for "D" and "F" schools respectively.

While these response rates dominate those achieved by the US Department of Education in the *Schools and Staffing Surveys* conducted in Florida at the same time as our surveys, one might still worry whether survey nonresponse might bias any of our results. While no test is perfect, we offer five pieces of evidence to suggest

[31] We emphasize that we do not necessarily believe that the grouped outcomes reflect one underlying latent construct. Rather, these are similar policies that schools might adopt.

[32] While a simple index is slightly more straightforward to understand, it also reweights individual outcomes within a domain when there is item nonresponse. Given the missing values in our data, we put more weight on the SUR estimates. That said, as shown in Table 6 (columns 4 and 5), the two methods give quite similar estimates.

[33] We excluded "alternative schools" such as adult schools, vocational/area voc-tech centers, schools administered by the Department of Juvenile Justice, and "other types" of schools. Note that we included charter schools serving "regular" students as well. The survey instruments are available upon request.

[34] The results are qualitatively similar when we control for the 1999–2000 survey responses in our models.

that survey nonresponse is not consequential for our application. First, while "D" and "F" schools have modestly lower response rates than do higher-rated schools, the response rate is not statistically different between the "F" and "D" schools. Second, as we show in Appendix Table A1, the characteristics of schools responding to the survey in 2003–2004, by school grade, are quite similar to those of nonrespondents. Third, we have estimated a series of regression-discontinuity models in which the dependent variable is either survey response in 2001–2002, survey response in 2003–2004, or survey response in 2003–2004, conditional on response in 2001–2002. In no instance, regardless of whether we control for grade points linearly, as quadratic, or as a cubic, was the indicator for being an "F" statistically significant at any conventional level. Fourth, we have regressed response status in 2003–2004 against 2003–2004 test scores and our set of controls and observe no relationship between test score performance and response status, conditioning on other covariates. Finally, we have constructed bounds akin to those employed by Angrist, Bettinger, and Kremer (2006) and Krueger (1999), to determine the degree to which differential attrition between our 2001–2002 and 2003–2004 survey rounds drive our results; the results of this exercise do not change our inference.[35]

We were careful to conduct the surveys in the early spring before schools knew which grade their school would receive for that academic year. Thus, when we conducted the survey in the spring of 2002, the administrators did not yet know their school grade (as school grades were announced on June 12, 2002, after the end of the school year and after our survey data collection had ceased.) Thus, we treat 2001–2002 as the "base year" and observe school policy as of the academic year of 2003–2004.[36]

IV. Results

A. *What Schools Were Doing in 2002*

We begin by presenting the policy/production environment of schools in 2002, the base year in our analysis. In Table 4, we report the means of the individual policies organized by their respective "domains" and by the school's grade in 2002. Among the variables that schools identified as strategies they employ to "improve low-performing students"[37] most of the schools required grade retention,[38] in-school supplemental instruction, and tutoring. Slightly less commonly schools required that students attend summer school and a few even required that such students

[35] In this bounding exercise, we assign all attriting schools a score equal to the minimum observed value in the domain or a score equal to the maximum observed value in the domain. No matter what we assume for missing values in 2003–2004, our results are fundamentally unchanged.

[36] Note that we do not observe school policies in the first year after the change in grading procedure; that is, in the 2002–2003 school year. We believe that not controlling for grades in 2003 likely biases the results downward since schools that received a grade of "F" in summer 2003 may have also changed their educational practices by the 2003–2004 school year. As shown in Table 1, 31 schools received an "F" in summer 2003 (18 elementary schools received an "F"). Seven of these schools had received an "F" in summer 2002, but no elementary schools received an "F" in both years.

[37] These responses are based on the question "What special measures, if any, does this school take to try to *improve* the performance of low-performing students?"

[38] Beginning in 2003 Florida required schools to retain all third graders that did not meet a predetermined threshold on the FCAT Sunshine State Standards reading test.

Table 4—2001–2002 School Response Variable Means

Domain/variable	School grade in 2002		
	A/B/C	D	F
Policies to improve low-performing students			
Require grade retention	0.76	0.82	0.79
Require summer school	0.40	0.57	0.36
Require before/after-school tutoring	0.44	0.62	0.68
Require in-school supplemental instruction	0.79	0.86	0.89
Require tutoring	0.61	0.72	0.82
Require Saturday classes	0.05	0.14	0.14
Require other policy	0.30	0.29	0.50
Lengthening instructional time			
Sponsor summer school	0.53	0.60	0.67
Sponsor year-round classes	0.01	0.01	0.00
Sponsor extended school year	0.19	0.24	0.25
Sponsor Saturday school	0.10	0.21	0.31
Sponsor after-school tutoring	0.78	0.88	0.86
Sponsor other school services	0.33	0.27	0.41
Average length of school day first and fourth grade (in minutes)	376.41	376.09	382.31
Reduced class size for subject			
Math	0.23	0.27	0.43
Reading	0.43	0.56	0.61
Writing	0.28	0.40	0.36
Low academic performance	0.44	0.55	0.68
Narrowing of curriculum			
Minimum time spent on math	0.67	0.81	0.86
Minimum time spent on reading	0.71	0.87	0.86
Minimum time spent on writing	0.62	0.81	0.75
Minimum time spent on social studies	0.43	0.45	0.61
Minimum time spent on art/music	0.59	0.64	0.71
Scheduling systems			
Block scheduling	0.35	0.43	0.52
Common prep periods	0.90	0.92	0.92
Subject matter specialist teachers	0.59	0.75	0.85
Organize teachers into teams	0.95	0.95	0.96
Looping students with same teacher in multiple years	0.43	0.41	0.33
Multi-age classrooms	0.29	0.37	0.46
Other schedule structure	0.11	0.07	0.15
Policies to improve low-performing teachers			
Supervise teachers more closely	0.98	0.99	1.00
Assign aide to teachers	0.30	0.52	0.59
Assign mentor to teachers	0.89	0.87	0.92
Provide additional professional development	0.99	1.00	1.00
Provide development/improvement plan	0.97	0.96	1.00
Other improvement strategy	0.14	0.13	0.20
Teacher resources			
Minutes per week for collaborative planning/class preparation	450.12	452.75	424.09
Days per year for individual professional development	3.24	3.94	5.08
Funds per student per year for professional development	$14.71	$28.48	$45.53
Teacher incentives			
Monetary reward (including one-time cash bonus)	0.29	0.22	0.29
Comp/release time	0.56	0.56	0.71
Choice of class	0.17	0.20	0.30
Attendance at conferences and workshops	0.64	0.65	0.71
Special leadership position/assignment	0.63	0.67	0.85
Other incentives	0.25	0.25	0.44

(*Continued*)

TABLE 4—2001–2002 SCHOOL RESPONSE VARIABLE MEANS (*Continued*)

Domain/variable	School grade in 2002		
	A/B/C	D	F
Teacher control (1 = *no influence*/5 = *complete control*)			
Teacher control of establishing curriculum	3.39	3.40	3.33
Teacher control of hiring new full-time teachers	2.90	2.75	3.00
Teacher control of budget spending	3.20	2.97	3.15
Teacher control of teacher evaluation	1.84	1.74	2.07
District control (1 = *no influence*/5 = *complete control*)			
District control of establishing curriculum	3.62	3.49	3.56
District control of hiring new full-time teachers	2.72	2.80	2.52
District control of budget spending	3.23	3.45	2.89
District control of teacher evaluation	2.28	2.41	2.19
Principal control (1 = *no influence*/5 = *complete control*)			
Principal control of establishing curriculum	3.54	3.51	3.48
Principal control of hiring new full-time teachers	4.35	4.26	4.39
Principal control of budget spending	3.90	3.77	4.00
Principal control of teacher evaluation	4.64	4.65	4.79
School climate (1 = *not at all accurate*/5 = *very accurate*)			
Staff morale is low (reversed)	4.18	3.91	3.93
Staff support and encourage each other	4.22	4.04	4.32
Teachers understand what is expected of them	4.40	4.14	4.57
New teachers (three or fewer years) are excellent	3.89	3.62	3.61
Experienced teachers (more than ten years) are excellent	4.07	3.74	3.69
Student disruption interferes with learning (reversed)	4.07	3.21	2.75
Parents worry about violence in school (reversed)	4.38	4.11	3.93
Parents closely monitor academic progress of child	3.29	2.68	2.46
Miscellaneous			
Reduced class size gifted academic performance	0.42	0.40	0.21
Average first and fourth grade class size	25.21	25.06	23.83
Minimum time spent on science	0.46	0.53	0.61
Whole school reform model	0.27	0.41	0.64

Notes: All variables within the "School Climate" domain have been coded to reflect a more positive school climate. For example, "staff morale is low" has been reversed such that a high value reflects the statement is "not at all accurate."

attend Saturday classes, although this policy was more prevalent among the "D"- and "F"-graded schools. Most schools also sponsored (although did not necessarily require) summer school and after-school tutoring. In addition, "F"-graded schools offered "other" school services such as extended day, mentors, or remedial instruction. Interestingly, the average class day was 376 minutes (or about 6 hours and 20 minutes), although it was slightly longer for the "F"-graded schools. Further, the average class size (not pupil-teacher ratio) in first and fourth grades was 25 students, with lower-graded schools having slightly smaller classes.

One strategy that some schools employ to improve achievement is to assign students to smaller "class units" for particular reasons. This does not mean that overall class sizes in a school are reduced, but they may, for instance, reorganize students either within classrooms or possibly even across classrooms for particular subjects. As an example, in our sample 61 percent of the "F"-graded schools used these smaller class "units" for reading compared to just over 40 percent of "A"-graded

schools. Similarly, "F"-graded schools were more likely to employ this strategy for math, writing, and low-performing students as well.

While the previous test score-based literature (including earlier versions of this paper) presents indirect evidence of "teaching to the test" as a response to accountability pressure, we also sought more direct evidence that might suggest this shift in focus, or at a minimum, a narrowing of the curriculum. We therefore asked the principals if the school had a policy on the minimum amount of time that fourth-grade students were required to spend on particular subjects (and, among those that did, what that minimum was). In 2002, 86 percent of the "F"-graded schools had policies that specified a minimum amount of time on math and reading, and three-quarters had a specified minimum amount of time on writing; these are all tested subjects. In contrast, only 61 percent had a policy on the minimum time spent on social studies, a subject that is not tested under Florida's state accountability system. The higher-graded schools were less likely to have school policies on how time was to be divided between the subjects.[39]

In terms of how schools attempt to schedule the day and organize how teachers work together, we observe that nearly all schools attempted to organize teachers into teams or tried to schedule common prep periods to provide teachers with an opportunity to collaborate. Similarly, a majority also employed subject-matter specialist teachers to assist other teachers. Finally, 15 percent of "F"-graded schools (and fewer of the higher-graded schools) employed some other way of structuring schedules and staff. Examples of these other strategies include: a school-wide common reading or language arts block (for example, 90 minutes of uninterrupted reading each day),[40] ability grouping, organizing even lower grades into "departments," or employing technology specialists to help teachers.

Given their centrality to the learning process, it is not surprising that schools employ a variety of strategies to improve low-performing teachers. As of the spring of 2002, nearly all schools supervised such teachers more closely, assigned an aide or mentor to them, and provided additional professional development and/or an improvement plan. However, in terms of the resources available to teachers, we find that the "F"-graded schools provided less time for collaborative planning and class preparation than higher-graded schools. They did, however, require more days per year for professional development, and had more money available for professional development activities each year (and more per teacher). Finally, while schools provided rewards for teacher performance (independent of any incentives used by the school district), the vast majority used comp or release time, attendance at conferences and workshops, or giving the teachers special leadership positions or assignments; fewer than one-third were found to use monetary rewards (including one-time cash bonuses).

[39] We collapse all of the "minimum time" questions together into the same domain, despite the fact that this combines the tested and untested subjects. Our construction of the "narrowing the curriculum" domain measure takes into account the fact that over-time changes in minimum time spent on tested subjects are negatively correlated with changes over time in the minimum time spent on untested subjects. We have also estimated models in which we limit the analysis to just tested subjects, and the results are similar to those presented herein: "F"-rated schools concentrate somewhat more on tested subjects as compared with "D"-rated schools.

[40] Implementing a common reading block is one of the corrective actions commonly adopted, according to the *Assistance Plus* reports.

We also asked a number of questions about the level of control that teachers, principals, and superintendents had over important decisions at the school (1 suggests no influence, whereas 5 suggests complete control). Across the board, principals had the most say in the hiring of new full-time teachers, how the school's budget would be spent, and over the evaluation of teachers. The school district superintendents had the most control over the curriculum.

Finally, while not necessarily the direct result of school policy, we were interested in gauging the climate at the school. The response indicates the extent to which the principal thought the statement was not at all accurate ("1") or was accurate ("5") of his or her school. We observe that in higher-graded schools the staff morale was higher, the teachers—new and experienced—were judged excellent, student disruption did not interfere with learning, violence was not a concern among parents, and parents closely monitored the academic progress of their children.

The fact that low-performing schools were already trying many of the policies and practices that our expert panel suggested underscores our point that there exists very little systematic evidence regarding what educational policies and practices actually work to improve student outcomes. This paper is about whether accountability pressure induces schools to try different things. Therefore, given this portrait of school policy and the environment as of the spring of 2002, the question is whether schools, the low-performing schools in particular, changed any of these policies or practices in response to their assigned performance grade.

B. *Changes Between 2002 and 2004*

Table 5 presents descriptive statistics of the change in school policy between the springs of 2002 and 2004 by the school's grade in the 2001–2002 school year.[41] Here, again, the individual variables are organized by their "domain." The "F"-graded schools appear to have been more likely to require grade retention and other forms of supplemental instruction for low-performing students, at least compared to higher-graded schools. Similarly, they were more likely to adopt summer school classes, extend the school year, and sponsor after-school tutoring. In addition, while they lengthened the school day by, on average, about two minutes, the higher-rated schools shortened the length of the school day. Although the "F"-schools were not more likely to institute minimum time requirements in the tested subjects, they were more likely to relax minimum time requirements on the untested subjects—such as social studies and art—a finding consistent with an incentive to narrow the curriculum in response to the accountability system.

In terms of teachers, the "F"-graded schools increased the amount of time available for collaborative planning and class preparation by about 80 minutes, compared

[41] One concern is that policies may appear to have changed because different people with different levels of information about what was going on in a school completed our survey. Unfortunately, we cannot readily assess in a systematic manner who completed the survey in each year. We have, however, pulled 100 random school surveys from 2001–2002 and compared the name of the person completing the survey in 2001–2002 to the name of the person completing the survey in 2003–2004. Out of 100 surveys, 88 had the same respondent in the two rounds and 97 had the same position title. This informal finding, combined with Chiang's (2008) result that school grades did not influence principal turnover, makes us less concerned that the findings are due to different people completing the survey.

Table 5—Mean Change in Variables between 2001–2002 and 2003–2004

	School grade in summer 2002		
Domain/variable	A/B/C	D	F
Policies to improve low-performing students			
Require grade retention	0.09	0.03	0.26
Require summer school	0.01	−0.11	−0.05
Require before/after-school tutoring	0.06	−0.04	0.15
Require in-school supplemental instruction	0.05	0.07	0.11
Require tutoring	0.07	0.00	0.10
Require Saturday classes	0.02	0.07	0.11
Require other policy	−0.02	0.04	−0.43
Lengthening instructional time			
Sponsor summer school	−0.03	0.00	0.10
Sponsor year-round classes (÷10)	−0.001	0.00	0.00
Sponsor extended school year	0.09	0.01	0.12
Sponsor Saturday school	0.04	0.09	0.05
Sponsor after-school tutoring	0.06	0.03	0.15
Sponsor other school services	0.06	−0.23	0.11
Average length of school day first and fourth grade (in minutes)	−0.98	−1.10	1.77
Reduced class size for subject			
Math	−0.01	−0.04	−0.05
Reading	−0.04	−0.10	−0.10
Writing	−0.05	−0.17	0.11
Low academic performance	0.02	−0.10	−0.21
Narrowing of curriculum			
Minimum time spent on math	0.08	0.07	0.05
Minimum time spent on reading	0.14	0.06	0.05
Minimum time spent on writing	0.02	−0.11	0.00
Minimum time spent on social studies	0.04	0.13	−0.25
Minimum time spent on art/music	0.03	0.13	−0.10
Scheduling systems			
Block scheduling	−0.01	0.00	0.18
Common prep periods	0.01	−0.01	0.00
Subject matter specialist teachers	0.01	0.06	0.06
Organize teachers into teams	0.01	−0.01	0.00
Looping (÷10)	0.01	−0.14	2.94
Multi-age classrooms (÷10)	0.03	−0.92	−1.87
Other schedule structure	−0.03	−0.04	0.00
Policies to improve low−performing teachers			
Supervise teachers more closely (÷10)	0.03	−0.14	0.00
Assign aide to teachers	−0.08	−0.10	−0.06
Assign mentor to teachers	0.02	0.13	0.05
Provide additional professional development (÷10)	−0.03	0.00	0.00
Provide development/improvement plan (÷10)	0.04	−0.14	0.00
Other improvement strategy	0.03	0.06	0.00
Teacher resources			
Minutes per week for collaborative planning/class preparation	11.10	−49.36	82.50
Days per year for individual professional development	0.49	−0.16	0.25
Funds per student per year for professional development	$33.13	$9.69	$42.33
Teacher incentives			
Monetary reward	0.03	0.07	0.06
Comp/release time (÷10)	−0.04	0.00	0.00
Choice of class	0.02	0.03	0.12
Attendance at conferences and workshops (÷10)	0.07	0.82	0.53
Special leadership position/assignment (÷10)	0.06	0.83	−1.11
Other incentives	−0.05	−0.08	0.13

(*Continued*)

TABLE 5—MEAN CHANGE IN VARIABLES BETWEEN 2001–2002 AND 2003–2004 (*Continued*)

Domain/variable	School grade in summer 2002		
	A/B/C	D	F
Teacher control (1 = *no influence*/5 = *complete control*)			
Teacher control of establishing curriculum	−0.11	−0.15	−0.10
Teacher control of hiring new full-time teachers	−0.02	−0.07	−0.20
Teacher control of budget spending	−0.08	−0.15	−0.05
Teacher control of teacher evaluation	−0.09	−0.04	−0.45
District/superintendent control (1 = *no influence*/5 = *complete control*)			
District control of establishing curriculum	0.14	0.18	0.20
District control of hiring new full-time teachers	0.00	−0.24	0.30
District control of budget spending	0.04	−0.29	0.44
District control of teacher evaluation	−0.01	−0.03	−0.30
Principal control (1 = *no influence*/5 = *complete control*)			
Principal control of establishing curriculum	−0.03	0.26	0.15
Principal control of hiring new full-time teachers	0.01	0.25	0.00
Principal control of budget spending	0.04	0.18	−0.11
Principal control of teacher evaluation	0.03	0.07	0.00
School climate (1 = *Not at all accurate*/5 = *Very accurate*)			
Staff morale is low (reversed)	0.08	0.10	0.21
Staff support and encourage each other	0.09	0.05	0.26
Teachers understand what is expected of them	0.05	0.09	−0.26
New teachers (three or fewer years) are excellent	−0.02	−0.07	0.00
Experienced teachers (more than ten years) are excellent	0.05	−0.15	0.17
Student disruption interferes with learning (reversed)	0.04	0.08	0.37
Parents worry about violence in school (reversed)	0.11	0.04	0.21
Parents closely monitor academic progress of child	0.02	−0.11	−0.32
Miscellaneous			
Reduced class size gifted academic performance	−0.01	−0.08	0.05
Average first and fourth grade class size	−2.06	−3.41	−2.70
Minimum time spent on science	0.05	0.13	−0.10
Whole school reform model	0.48	0.28	0.00

Notes: All variables within the "School Climate" domain have been coded to reflect a more positive school climate. For example, "staff morale is low" has been reversed such that a high value reflects the statement is "not at all accurate."

to only an 11-minute increase for the "A"-graded schools and a nearly one-hour decrease for the "D"-graded schools. In general, teachers lost some autonomy over important decisions made at the school.

Finally, it is notable that all schools appear to have reduced class sizes between the two years by between two to three students, particularly among the lower-rated schools.

While the above descriptive changes are suggestive of an association between a school receiving an "F" grade and the policies and practices adopted by the school to improve, they do not necessarily indicate causation. We attempt to discern the causal effect of receiving an "F" grade in the models whose results are presented in Table 6. (We present OLS results for selected individual variables in Table 7).[42]

[42] The results are similar if we cluster the standard errors at the district level. We also note that we do not report results adjusting the standard errors for multiple testing primarily because we do not expect that schools will

Table 6—Seemingly-Unrelated Regression and OLS Results of the Effect of Receiving an "F" Grade versus a "D" Grade in Summer 2002 on School Policy in 2003–2004

	Seemingly unrelated regression				Index
	F versus all comparison		F versus D comparison		
Domain	(1)	(2)	(3)	(4)	(5)
Policies to improve low-performing students	0.437	0.325	0.321	0.308	0.345
	(0.097)	(0.128)	(0.128)	(0.139)	(0.167)
Lengthening instructional time	0.279	0.283	0.280	0.264	0.237
	(0.095)	(0.122)	(0.123)	(0.123)	(0.139)
Reduced class size for subject	0.557	0.342	0.340	0.299	0.364
	(0.190)	(0.240)	(0.240)	(0.240)	(0.258)
Narrowing of curriculum	0.140	0.058	0.060	0.137	0.123
	(0.086)	(0.102)	(0.102)	(0.107)	(0.124)
Scheduling systems	0.208	0.255	0.249	0.377	0.361
	(0.120)	(0.144)	(0.145)	(0.139)	(0.143)
Policies to improve low-performing teachers	0.215	0.192	0.190	0.154	0.204
	(0.080)	(0.103)	(0.103)	(0.097)	(0.156)
Teacher resources	0.723	0.956	0.963	1.048	0.867
	(0.390)	(0.496)	(0.492)	(0.532)	(0.672)
Teacher incentives	0.220	0.125	0.124	0.125	0.170
	(0.160)	(0.202)	(0.202)	(0.202)	(0.202)
Teacher autonomy	−0.060	−0.026	−0.024	−0.091	−0.095
	(0.129)	(0.162)	(0.161)	(0.158)	(0.199)
District control	−0.169	−0.065	−0.069	0.082	0.066
	(0.172)	(0.203)	(0.202)	(0.203)	(0.201)
Principal control	0.169	−0.186	−0.189	−0.294	−0.329
	(0.129)	(0.159)	(0.159)	(0.161)	(0.206)
School climate	−0.454	0.165	0.165	0.121	0.118
	(0.124)	(0.158)	(0.158)	(0.157)	(0.160)
Specifications also control for:					
2002 grade points	N	Y	Y	Y	Y
2002 "Simulated Grade"	N	N	N	Y	Y
Lagged dependent variable	N	N	N	Y	Y
Other school characteristics	N	N	N	Y	Y

Notes: Standard errors are in parentheses. The estimates in columns 1−4 are based on seemingly-unrelated regressions of variables in each of the "domains" as described in Tables 4 and 5; the estimates in column 5 use an "index" (or mean standardized value) of the variables in each of the domains, as described in the text. The estimates in columns 1 and 2 compare the F-graded schools to all higher-grade schools; those in columns 3–5 compare the F-graded schools to the D-graded schools. "2002 grade points" is a cubic in the number of points the school received according to the summer 2002 accountability grading system. The "2002 Simulated Grade" is a vector of dummy variables indicating the grade the school would have received in 2002 using the 2001 school grading formula. The "lagged dependent variable" is the school's value for the 2004 policy in 2002 an indicator for whether the 2002 response is missing. "Other school characteristics" include: racial and ethnic composition of the school, the school's estimated expenditures per student in 2002 for special education, vocational education, education for "at-risk" students, and "regular" education, the number of students in the school, the racial and ethnic composition of the school, the percentage of the school's student body that is classified as disabled, eligible for free or reduced-price lunch, limited English proficient, and gifted, and the "stability rate" of the school.

The first four columns of Table 6 use SUR to aggregate the responses within each domain; the final column simply creates an "index" of the individual components.

adopt policies in all of the domains, and do not have a prior as to what fraction of outcomes should be significantly affected.

TABLE 7—OLS RESULTS OF THE EFFECT OF RECEIVING AN "F" GRADE VERSUS A "D" GRADE IN SUMMER 2002 ON SCHOOL SELECTED INDIVIDUAL POLICIES IN 2003–2004

	F versus all comparison		F versus D
Variable	No covariates (1)	With a cubic in grade points (2)	With all covariates (3)
Average first and fourth grade class size	−2.946	−1.276	−0.747
	(0.733)	(0.953)	(0.908)
Use block scheduling	0.250	0.095	0.247
	(0.097)	(0.124)	(0.115)
Use common prep period	0.010	0.113	0.111
	(0.054)	(0.070)	(0.071)
Use other scheduling structure	0.066	0.134	0.185
	(0.077)	(0.100)	(0.106)
Minutes per week for collaborative planning/class preparation	21.791	92.641	123.702
	(45.482)	(60.285)	(62.288)
Reduced class size gifted academic performance	−0.245	−0.141	−0.065
	(0.099)	(0.129)	(0.127)
Whole school reform model	−0.056	−0.026	−0.019
	(0.085)	(0.111)	(0.110)
Minimum time spent on science	0.101	−0.055	−0.127
	(0.097)	(0.128)	(0.119)

Notes: Standard errors are in parentheses. The estimates in columns 1 and 2 compare the F-graded schools to all higher-grade schools; those in column 3 compare the F-graded schools to the D-graded schools. The estimates in columns 2 and 3 control for a cubic in the number of points the school received according to the summer 2002 accountability grading system. Those in column 3 also control for the "2002 Simulated Grade" (a vector of dummy variables indicating the grade the school would have received in 2002 using the 2001 school grading formula); the "lagged dependent variable" (the school's value for the 2004 policy in 2002 and an indicator for whether the 2002 response is missing); and "Other school characteristics" (the school's estimated expenditures per student in 2002 for special education, vocational education, education for "at-risk" students, and "regular" education, the number of students in the school, the racial and ethnic composition of the school, the percentage of the school's student body that is classified as disabled, eligible for free or reduced-price lunch, limited English proficient, and gifted, and the "stability rate" of the school).

In the first two columns, we present the effect by comparing the "F"-graded schools to all higher graded schools; in columns 3–5 we compare the "F"-graded schools to "D"-graded schools.

The first column does not include any other covariates and we estimate that schools that received an "F"-grade were more likely to adopt policies to improve low-performing students, lengthened instruction time, assigned students to smaller class "units" for particular reasons, narrowed the curriculum, changed their scheduling systems, adopted policies to improve low-performing teachers, increased resources available to teachers, and had a worsening climate.

Column 2 employs a basic version of our regression discontinuity design in which we control for a cubic in the total number of grade points earned by the school in 2002. In just over one-half of the cases, the coefficient estimates decrease conditional on the grade points; in the other cases the coefficients increase or reverse sign. In column 3, we narrow the focus even further by comparing the effect of receiving an "F"-grade to that of receiving a "D"-grade while continuing to control for the cubic in the grade points (column 2). Note that the estimates are very similar to those in column 2.

As the validity of these estimates hinges on the regression discontinuity, we conducted two falsification checks that are not shown here but are available upon request. Specifically, we estimated whether the baseline school characteristics and/or school policies in 2002 (described in Section IVB) differed between "D" and "F" schools, conditional on a cubic in grade points. It appears that the "F" schools had higher proportions of black and Asian students, a lower stability rate, and were smaller than the "D" schools, conditional on the grade points; along other dimensions the two sets of schools look statistically similar. As for discontinuities among school policies in 2002, three policies were significant at the 10 percent level: those that would narrow the curriculum (although the sign was negative suggesting that the "F" schools were "broadening" compared to "D" schools in 2002), those that would provide incentives to teachers, and characteristics of a positive school climate.

These results raise the possibility that the point estimates in column 3 are biased. However, when we control for the school's 2002 "simulated grade"—thus taking into account whether or not the school "expected" to receive an "F" grade, along with the 2002 level of the 2004 policy; an indicator for a missing response in 2002); and other school characteristics (column 4)—our estimates largely remain similar.[43] One exception is that in column 4 the loss of control by the school principal is now statistically significant at the 10 percent level. Overall, while these estimates may be biased due to unobservable school characteristics, we suspect that it is unlikely given the relative surprise of the new grading system and the fact that the estimates are quite stable as we narrow the focus to the difference between "D" and "F" schools and as we control for many school characteristics and policies.

Finally, in column 5, we compare the results that rely on a SUR analysis to those generated by simply computing an index (or standardized mean value) of the associated items in the domain. Again, the results do not generally change much although they are less precise. We primarily show these estimates as they facilitate the analysis in the last section.

Table 7 shows OLS results of individual items within the domains that were statistically significant at the 10 percent level (or slightly above) as well as of a few items that we could not easily fit into a domain. In the first column we show the effect on "F"-graded schools compared to all schools and we do not control for any covariates. In the second column we continue to show the effect on "F"-graded schools compared to all schools, but also control for a cubic in grade points (similar to column 2 of Table 6), thereby adopting a regression discontinuity design. The third column is similar to column 4 of Table 6 in that it compares the "F"-graded schools to the "D"-graded schools, conditional on the full set of covariates.

[43] We have also included the lagged variables (and indicators for these variables being missing) for all of the items in the domain with similar results. We only report results controlling for the 2001–2002 outcome for parsimony. In addition, as noted earlier, we also obtain similar results if we also include the 1999–2000 level of the variable. We also include flags for missing expenditures or characteristics.

We find that schools facing the increased pressure adopt block scheduling, re-organize the school scheduling structure through "other" measures, and increase time for collaborative planning and class preparation for teachers; we also see some evidence that they provide a common prep period for teachers, and increase the district's control over how the budget is spent. And while it may appear that "F"-graded schools are more likely to reduce class sizes to improve student achievement, this response substantially decreases and becomes statistically insignificant when we add covariates.

In sum, we find that schools receiving an "F" grade are more likely to focus on low-performing students, lengthen the amount of time devoted to instruction, adopt different ways to organize the day and learning environment of the students and teachers, increase resources available to teachers, and decrease principal control, all of which is as might be expected given the increased oversight built into the A+ Plan.

C. *Are the Policies Related to the Change in Test Scores?*

Given that we find that the increased accountability pressure associated with receipt of an "F" grade led to substantial changes in instructional policies and practices, it is natural to ask whether these changes in policies and practices are associated with the improved test scores found by Chiang (2009), Rouse et al. (2007), and West and Peterson (2006). Unfortunately, we cannot obtain causal estimates of the effectiveness of the individual policies because we only have one instrument for the school policy and practice variables. We can, however, observe whether the inclusion of these policies and practices in a test score equation helps to explain the observed test score improvements associated with receipt of an "F" grade.

In order to fix comparisons, the first row of Table 8 reports the results of a test score regression discontinuity model first presented in an earlier version of this paper (Rouse et al. 2007).[44] We relate a student's seventh grade test score in 2004–2005 to the accountability status faced by the student's elementary school while he/she was in fifth grade (2002–2003), the first year of the new accountability regime, controlling for pre-fifth grade test scores and a variety of student characteristics.[45,46] We conduct this analysis in an attempt to identify lasting positive effects—rather than short-term transitory effects—of accountability pressure for student test scores.[47] We observe that the receipt of an "F" grade is associated with longer-term improvements of 8 or 9 percent of a standard deviation improvements

[44] Both Rouse et al. (2007) and Chiang (2009) include a wide variety of specification checks.

[45] Specifically, we control for a cubic in fourth grade norm-referenced test scores (as well as indicators for missed scores), grade dummies, year dummies, indicators for current and historical disability status, race, ethnicity, English language learner status, free lunch eligibility, reduced lunch eligibility, subsidized lunch ineligibility conditional on application, prior year's school grade, the simulated grade that would have occurred in 2002 had the school grading system not changed, a cubic in the number of points the school received according to the 2002 accountability grading system, and elementary-middle combination fixed effects. We use norm-referenced tests so that we can condition on a low-stakes test measure, but our results are fundamentally identical if we instead control for fourth-grade criterion-referenced tests.

[46] We find very similar results when we consider test scores in 2003–2004, which is when this cohort would be in sixth grade, the years when our other dependent variables are mentioned. We focus on seventh grade scores to consider somewhat longer-term outcomes and to make our analysis more parallel to Chiang (2009).

[47] Rouse et al. (2007) and Chiang (2009) both present results for shorter-term outcomes as well.

Table 8—The Effect of Including School Policy/Practice Variables on Regression-Discontinuity Estimates of the Effect of Receiving an "F" versus "D" Grade in Summer 2002 on Seventh-Grade Student Performance: Fifth-Grade Cohort of 2002–2003

Model specification	Standardized test score			
	High-stakes reading	High-stakes math	Low-stakes reading	Low-stakes math
Models excluding school policy/practice variables	0.080	0.094	0.059	0.045
	(0.030)	(0.039)	(0.025)	(0.027)
Models including all school policy/practice variables	0.068	0.058	0.042	0.022
	(0.037)	(0.042)	(0.027)	(0.028)
p-value of joint significance of policy variables	*0.002*	*0.001*	*0.001*	*0.007*
Models including only the five school policy/practice domains with the strongest estimated relationship with "F" grade status	0.068	0.053	0.044	0.011
	(0.036)	(0.041)	(0.026)	(0.030)
p-value of joint significance of policy variables	*0.054*	*0.041*	*0.012*	*0.042*

Notes: Standard errors adjusted for school-level clustering are in parentheses. Dependent variables are test scores standardized to the Florida average by grade level. Regressions control for a cubic in fourth grade norm-referenced test scores (as well as indicators for missed scores), grade dummies, year dummies, indicators for current and historical disability status, race, ethnicity, English language learner status, free-lunch eligibility, reduced-lunch eligibility, subsidized-lunch ineligibility conditional on application, prior year's school grade, the simulated grade that would have occurred in 2002 had the school grading system not changed, a cubic in the number of points the school received according to the 2002 accountability grading system, and elementary-middle combination fixed effects. The second set of regressions also includes the twelve policy indices described in Table 6, as well as the class size variable, which itself is individually statistically significantly related to F grade receipt in summer 2002. The third set of regressions includes the five domains with the highest estimated relationship with "F" grade status (see Table 6).

in reading and math scores on high-stakes exams, and 5 or 6 percent of a standard deviation improvements in reading and math scores on low-stakes exams. The point of this analysis, however, is to gauge the degree to which these results are affected when we control for the policies and practices identified in our surveys. Thus, the second and third sets of results reported in Table 8 are the "F" grade coefficients in models that also control for both the policy indices presented in Table 6, and the class size variable described in Table 7 (or those that include just the five domains that we found to have the strongest estimated relationship with "F" grade status, as shown in Table 6).[48]

In the table we present *p*-values of tests of joint significance, depending on model specification, either of the five domains most highly related to "F" grade status; or (in the case of the model in which we include all policy domains) of the full set of policy variables considered. We find that in all specifications, the set of policy variables that we control for—whether the full set of policy domains or the set of domains most related to "F" grade status—are jointly significant.

More to the point of the analysis is the percentage reduction in the estimated "F"-grade test score effects when these policy variables are included as control variables. Across the specifications (including in other grades not reported in the table), the estimated effect of "F" grade receipt decreases with the inclusion of these policy

[48] We present evidence for the four different test scores for seventh graders, conditional on lagged fourth grade test scores, although the results are comparable when we look instead at fifth graders or sixth graders.

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 44 percent with regard to mathematics. Moreover, virtually all the 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 relationship with "F" grade receipt. While we do not ascribe a causal interpretation to these findings, the results appear to indicate that the differential test score gains between "F"-graded schools and "D"-graded schools appear to be related to the different ways in which these schools implemented school policy variables captured in our surveys. It is, however, impossible to know which of these variables are responsible for the test score gains, as any exercise in ascertaining causality would require a very different identification strategy.

V. Conclusion

This project presents an attempt to systematically get inside the "black box" of the effects of school accountability systems on student performance. In this paper, we find that schools that received a grade of "F" in summer 2002 engaged in systematically different changes in instructional policies and practices as a consequence of school accountability pressure, and that these policy changes explain a significant share of the test score improvements (in some subject areas) associated with "F"-grade receipt.

We note also that the remaining estimated "F"-grade effect may be due to other unobserved changes in policies or programs, or additionally, to an increase in the productivity of school resources due to the "F"-label (e.g., through an increase in effort on the part of the school superintendent, principals, and teachers). Further, we observe that the "F" schools adopted different policies than did "D" (or higher-graded) schools. And while we think that these policies reflect decisions made by the school principal and/or superintendent in an attempt to improve school productivity, we do not strictly observe who may have instigated them, and under the *Assistance Plus* program there is scope for others to have much influence.

Our findings are provocative in that they suggest that there are some potential policies and practices that low-performing schools may have used to improve their performance when they faced increased accountability pressure. However, while suggesting that school accountability systems structured similarly to the A+ Plan *can* spur improvement, it is premature to outline a prescription for the improvement of low-performing schools based on these findings, particularly since we do not observe student performance along all relevant dimensions. A comprehensive study of school efficiency would also include untested domains, such as student performance in social studies and the arts. And, of course, our instrument for accountability pressure is the same for all of our policy and practice domains, so we cannot credibly identify any specific pathways through which the accountability pressure is working. That said, we find that accountability pressures can induce school administrators to change their behavior in educationally beneficial ways that collectively seem to have a meaningful effect on student outcomes.

Appendix

Table A1—Difference in School Characteristics between Respondents and Nonrespondents to the 2004 School Survey Sample, by the School's Accountability Grade in 2001–2002

	School grade in 2002		
Variable	A/B/C	D	F
Number of students	1.356	39.510	−74.378
	(15.956)	(41.448)	(86.885)
Proportion black	−0.041	−0.055	−0.085
	(0.014)	(0.055)	(0.085)
Proportion Asian	0.000	0.005	−0.002
	(0.001)	(0.003)	(0.002)
Proportion Hispanic	−0.029	−0.028	0.021
	(0.014)	(0.039)	(0.060)
Proportion Native American	0.000	−0.001	0.000
	(0.000)	(0.001)	(0.001)
Proportion mixed race	−0.002	0.002	0.000
	(0.001)	(0.002)	(0.004)
Proportion free or reduced-price lunch eligible	−1.902	1.176	−0.622
	(1.453)	(2.787)	(6.072)
Proportion gifted	0.222	0.062	−0.421
	(0.337)	(0.360)	(0.568)
Proportion limited English proficiency	−1.404	−1.664	1.222
	(0.704)	(2.169)	(3.983)
Proportion classified disabled	1.255	0.499	3.159
	(0.351)	(1.206)	(2.157)
Stability rate	0.144	−0.071	2.899
	(0.177)	(0.594)	(1.215)
Special education expenditures per pupil	−$168.387	−$28.921	$575.636
	(131.352)	(430.789)	(757.596)
Regular education expenditures per pupil	−$118.849	−$313.196	$528.114
	(52.823)	(228.127)	(394.682)
At-risk students expenditures per pupil	−$119.886	−$2.533	$1,421.659
	(236.349)	(741.812)	(1,521.961)
Vocational education expenditures per pupil	−$4.938	−$61.475	−$1,318.667
	(51.924)	(257.009)	(961.135)

Notes: Standard errors in parentheses. All school characteristics are based on data from the 2001–2002 school year. The data include elementary schools in both the 2002 and 2004 sampling frames.

REFERENCES

Akerlof, George A., and Rachel E. Kranton. 2005. "Identity and the Economics of Organizations." *Journal of Economic Perspectives* 19 (1): 9–32.

Akerlof, George A., and Rachel E. Kranton. 2007. "More than Money: Economics and Identity." Unpublished.

Anderson, Patricia, Kristin Butcher, and Diane Whitmore Schanzenbach. 2011. "Adequate (or Adipose?) Yearly Progress: Assessing the Effect of 'No Child Left Behind' on Children's Obesity." National Bureau of Economic Research (NBER) Working Paper 16873.

Angrist, Joshua D., Eric Bettinger, and Michael Kremer. 2006. "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *American Economic Review* 96 (3): 847–62.

Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics* 114 (2): 533–75.

Bacolod, Marigee, John DiNardo, and Mireille Jacobson. 2009. "Beyond Incentives: Do Schools Use Accountability Rewards Productively?" National Bureau of Economic Research (NBER) Working Paper 14775.

Ballou, Dale, and Matthew Springer. 2008. "Achievement Tradeoffs and No Child Left Behind." Unpublished.

Booher-Jennings, Jennifer. 2005. "Below the Bubble: 'Educational Triage' and the Texas Accountability System." *American Educational Research Journal* 42 (2): 231–68.

Carnoy, Martin, and Susanna Loeb. 2002. "Does External Accountability Affect Student Outcomes? A Cross State Analysis." *Education Evaluation and Policy Analysis* 24 (4): 305–31.

Chakrabarti, Rajashri. 2007. "Vouchers, Public School Response and the Role of Incentives: Evidence from Florida." Federal Reserve Bank (FRB) of New York Working Paper 306.

Chiang, Hanley. 2008. "How Accountability Pressure on Failing Schools Affects Student Achievement." Mathematica Policy Research Working Paper 6364.

Chiang, Hanley. 2009. "How Accountability Pressure on Failing Schools Affects Student Achievement." *Journal of Public Economics* 93 (9–10): 1045–57.

Chubb, John, and Terry Moe. 1990. *Politics, Markets and America's Schools*. Washington, DC: Brookings Institution.

Clark, Melissa A. 2002. "Education Reform, Redistribution, and Student Achievement: Evidence from the Kentucky Education Reform Act." Princeton University Working Paper November.

Craig, Steven, Scott Imberman, and Adam Perdue. 2013. "Does it Pay to Get an A? School Resource Allocation in Response to Accountability Ratings." *Journal of Urban Economics* 73 (1): 30–42.

Cullen, Julie Berry, and Michael J. Mazzeo. 2007. "Implicit Performance Awards: An Empirical Analysis of the Labor Market for Public School Administrators." Unpublished.

Cullen, Julie Berry, and Randall Reback. 2006. "Tinkering Toward Accolades: School Gaming Under a Performance Accountability System." In *Improving School Accountability: Check-Ups or Choice, Advances in Applied Microeconomics*, Vol. 14, edited by Timothy J. Gronberg and Dennis W. Jansen, 1–34. Amsterdam: Elsevier Science.

Dee, Thomas S., and Brian Jacob. 2011. "The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management* 30 (3): 418–46.

Feng, Li, David Figlio, and Tim Sass. 2010. "School Accountability and Teacher Mobility." National Bureau of Economic Research (NBER) Working Paper 16070.

Figlio, David N. 2006. "Testing, Crime and Punishment." *Journal of Public Economics* 90 (4–5): 837–51.

Figlio, David N., and Lawrence Getzler. 2006. "Accountability, Ability and Disability: Gaming the System?" In *Improving School Accountability: Check-Ups or Choice, Advances in Applied Microeconomics*, Vol. 14, edited by Timothy J. Gronberg and Dennis W. Jansen, 35–50. Amsterdam: Elsevier Science.

Figlio, David N., and Lawrence W. Kenny. 2009. "Public Sector Performance Measurement and Stakeholder Support." *Journal of Public Economics* 93 (9–10): 1069–77.

Figlio, David, and Susanna Loeb. 2010. "School Accountability." In *Handbook of the Economics of Education*, Vol. 3, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 383–422. Amsterdam: North-Holland.

Figlio, David N., and Maurice E. Lucas. 2004. "What's in a Grade? School Report Cards and the Housing Market." *American Economic Review* 94 (3): 591–604.

Figlio, David N., and Cecilia Elena Rouse. 2006. "Do Accountability and Voucher Threats Improve Low-Performing Schools?" *Journal of Public Economics* 90 (1–2): 239–55.

Figlio, David N., and Joshua Winicki. 2005. "Food for Thought: The Effects of School Accountability Plans on School Nutrition." *Journal of Public Economics* 89 (2–3): 381–94.

Goldhaber, Dan. 2009. "Voucher Finance." In *Handbook of Research on School Choice*, edited by Mark Berends, Matthew G. Springer, Dale Ballou, and Herbert J. Wahlberg, 309–20. New York: Routledge.

Goldhaber, Dan, and Jane Hannaway. 2004. "Accountability with a Kicker: Preliminary Observations on the Florida A+ Accountability Plan." *Phi Delta Kappan* 85 (8): 598–605.

Greene, Jay, Julie Trivitt, and Marcus Winters. 2009. "The Impact of High-Stakes Testing on Student Proficiency in Low-Stakes Subjects: Evidence from Florida's Elementary Science Exam." *Economics of Education Review* 29 (1): 138–46.

Hamilton, Laura, Brian M. Stecher, Julie A. Marsh, Jennifer Solan McCombs, Abby Robyn, Jennifer Russell, Scott Naftel, and Heather Barney. 2007. *Standards-Based Accountability under No Child Left Behind: Experiences of Teachers and Administrators in Three States*. Santa Monica: RAND Corporation.

Haney, Walt. 2000. "The Myth of the Texas Miracle in Education." *Education Policy Analysis Archives* 8 (41).

Haney, Walt. 2002. "Lake Woebe Guaranteed: Misuse of Test Scores in Massachusetts, Part I." *Education Policy Analysis Archives* 10 (24).

Hannaway, Jane, and Laura Hamilton. 2007. *Effects of Accountability Policies on Classroom Practices*. Washington, DC: Urban Institute.

Hannaway, Jane, and Kristi Kimball. 2001. "Big Isn't Always Bad: School District Size, Poverty, and Standards-Based Reform." In *From the Capitol to the Classroom: Standards-Based Reform in the States*, edited by Susan Fuhrman, 99–123. Chicago: National Society for the Study of Education.

Hanushek, Eric A., Charles S. Benson, Richard B. Freeman, Dean T. Jamison, Henry M. Levin, Rebecca A. Maynard, and Richard J. Murnane, et al. 1994. *Making Schools Work: Improving Performance and Controlling Costs*. Washington, DC: Brookings Institution.

Hanushek, Eric A., and Dale W. Jorgenson, eds. 1996. *Improving America's Schools: The Role of Incentives*. Washington, DC: National Academy Press.

Hanushek, Eric A., and Margaret E. Raymond. 2005. "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management* 24 (2): 297–327.

Hout, Michael, and Stuart W. Elliott, eds. 2011. *Incentives and Test-Based Accountability in Education*. National Research Council of National Academies. Washington, DC, May.

Howell, William, Paul Peterson, Patrick Wolf, and David Campbell. 2002. *The Education Gap: Vouchers and Urban Schools*. Washington, DC: Brookings Institution.

Jacob, Brian A. 2005. "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics* 89 (5–6): 761–96.

Jacob, Brian A., and Lars Lefgren. 2004. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." *Review of Economics and Statistics* 86 (1): 226–44.

Jacob, Brian A., and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics* 118 (3): 843–77.

Kling, Jeffrey R., and Jeffrey B. Liebman. 2004. "Experimental Analyses of Neighborhood Effects on Youth." Princeton University Industrial Relations Section Working Paper 483.

Koretz, Daniel M. 2003. "Using Multiple Measures to Address Perverse Incentives and Score Inflation." *Educational Measurement: Issues and Practice* 22 (2): 18–26.

Koretz, Daniel M., and Sheila I. Barron. 1998. *The Validity of Gains in Scores on the Kentucky Instructional Results Information System* (*KIRIS*). Research and Development Corporation (RAND). Santa Monica.

Krieg, John. 2008. "Are Students Left Behind? The Distributional Effects of the No Child Left Behind Act." *Education Finance and Policy* 3 (2): 250–81.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114 (2): 497–532.

Krueger, Alan B., and Diane M. Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR." *Economic Journal* 111 (468): 1–28.

Krueger, Alan B., and Pei Zhu. 2004. "Another Look at the New York City Voucher Experiment." *American Behavioral Scientist* 47 (5): 658–98.

Ladd, Helen, and Douglas Lauen. 2009. "Status Versus Growth: The Distributional Effects of School Accountability Policies." Urban Institute Working Paper #21.

Lee, Jaekyung. 2008. "Is Test-Driven External Accountability Effective? Synthesizing the Evidence from Cross-State Causal-Comparative and Correlational Studies." *Review of Educational Research* 8 (3): 608–44.

Linn, Robert. 2005. "Alignment, High Stakes, and the Inflation of Test Scores." In *Uses and Misuses of Data for Educational and Accountability Improvement*, edited by Joan L. Herman and Edward H. Haertel, 99–118. Malden: Blackwell Publishing.

Neal, Derek, and Diane Whitmore Schanzenbach. 2010. "Left behind by Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics* 92 (2): 263–83.

Ozek, Umut. 2010. "One Day Too Late? Mobile Students in an Era of Accountability." Unpublished.

Reback, Randall. 2008. "Teaching to the Rating: School Accountability and the Distribution of Student Achievement." *Journal of Public Economics* 92 (5–6): 1394–1415.

Reback, Randall, Jonah Rockoff, and Heather Schwartz. 2011. "Under Pressure: Job Security, Resource Allocation, and Productivity in Schools under NCLB." National Bureau of Economic Research (NBER) Working Paper 16745.

Ringwalt, Chris, Sean Hanley, Susan Ennett, Amy Vincus, J. Michael Bowling, Susan Haws, and Louise Rohrbach. 2011. "The Effects of No Child Left Behind on the Prevalence of Evidence-Based Drug Prevention Curricula in the Nation's Middle Schools." *Journal of School Health* 81 (5): 265–72.

Rockoff, Jonah, and Lesley J. Turner. 2010. "Short-Run Impacts of Accountability on School Quality." *American Economic Journal: Economic Policy* 2 (4): 119–47.

Rouse, Cecilia Elena. 1998. "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program." *Quarterly Journal of Economics* 113 (2): 553–602.

Rouse, Cecilia Elena, Jane Hannaway, Dan Goldhaber, and David Figlio. 2007. "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure." National Bureau of Economic Research (NBER) Working Paper 13681.

Rouse, Cecilia Elena, Jane Hannaway, Dan Goldhaber, and David Figlio. 2013. "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure: Dataset." *American Economic Journal: Economic Policy*. http://dx.doi.org/10.1257/pol.5.2.251.

Simon, Herbert A. 1957. *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations*. 2nd ed. New York: Free Press.

Smith, Marshall S., and Jennifer A. O'Day. 1991. "Systemic School Reform." In *The Politics of Curriculum and Testing: The 1990 Yearbook of the Politics of Education Association*, edited by Susuan H. Fuhman and Betty Malen, 233–67. New York: Falmer Press.

Stecher, Brian. 2002. "Consequences of Large-Scale, High-Stakes Testing on School and Classroom Practice." In *Making Sense of Test-Based Accountability in Education*, edited by Laura S. Hamilton, Brian M. Stecher, and Stephen P. Klein, 79–100. Santa Monica: RAND Corporation.

West, Martin R., and Paul E. Peterson. 2006. "The Efficacy of Choice Threats within School Accountability Systems: Results from Legislatively Induced Experiments." *Economic Journal* 116 (510): C46–62.

White, Katie, and James Rosenbaum. 2008. "Inside the Black Box of Accountability: How High-Stakes Accountability Alters School Culture and the Classification and Treatment of Students and Teachers." In *No Child Left Behind and the Reduction of the Achievement Gap: Sociological Perspectives on Federal Education Policy*, edited by Alan R. Sadovnik, Jennifer A. O'Day, George W. Bohrnstedt, and Kathryn M. Borman, 97–116. New York: Routledge.

Wong, Manyee, Thomas Cook, and Peter Steiner. 2010. "No Child Left Behind: An Interim Evaluation of Its Effects on Learning Using Two Interrupted Time Series Each with Its Own Non-Equivalent Comparison Series." Northwestern University Working Paper 09-11.

Yin, Lu. 2011. "Are School Accountability Systems Contributing to Adolescent Obesity?" Unpublished.

This article has been cited by:

1. Atila Abdulkadiroğlu, Parag A. Pathak, Christopher R. Walters. 2018. Free to Choose: Can School Choice Reduce Student Achievement?. *American Economic Journal: Applied Economics* **10**:1, 175-206. [Abstract] [View PDF article] [PDF with links]
2. Dennis Epple, Richard E. Romano, Miguel Urquiola. 2017. School Vouchers: A Survey of the Economics Literature. *Journal of Economic Literature* **55**:2, 441-492. [Abstract] [View PDF article] [PDF with links]
3. David J. Deming, David Figlio. 2016. Accountability in US Education: Applying Lessons from K–12 Experience to Higher Education. *Journal of Economic Perspectives* **30**:3, 33-56. [Abstract] [View PDF article] [PDF with links]
4. David Figlio, Cassandra M. D. Hart. 2014. Competitive Effects of Means-Tested School Vouchers. *American Economic Journal: Applied Economics* **6**:1, 133-156. [Abstract] [View PDF article] [PDF with links]