Leaving No Child Behind: Two Paths to School Accountability

David N. Figlio
University of Florida and NBER
figlio@northwestern.edu

Cecilia E. Rouse
Princeton University and NBER

Analia Schlosser
Princeton University and Tel Aviv University
analias@post.tau.ac.il

Working Paper

**Please do not quote or cite without permission*

Paper presented at the *NCLB: Emerging Findings Research Conference* at the Urban Institute, Washington, D.C. on August 12, 2009. The conference was supported, in part, by the National Center for the Analysis of Longitudinal Data in Education Research (CALDER), funded by Grant R305A060018 to the Urban Institute from the Institute of Education Sciences, U.S. Department of Education and the National Center for Performance Incentives (NCPI) at Vanderbilt University. The authors thank Lisa Markman for helpful conversations. They are grateful to the Florida Department of Education for providing student-level data necessary to conduct this research and indebted to the Annie E. Casey, Atlantic Philanthropies, Smith Richardson and Spencer Foundations, the U.S. Department of Education and the National Institutes of Health for financial support. The views expressed in the paper are solely those of the authors and may not reflect those of the funders or supporting organizations. Any errors are attributable to the authors.

Abstract

The relatively poor academic achievement of black and Hispanic students has been a national concern since the passage of the *Elementary Secondary and Education Act* in 1963. Frustrated with relatively slow progress in closing these educational gaps, the most recent reauthorization of the ESEA, the *No Children Left Behind Act of 2001* (NCLB) attempts to employ rigorous accountability standards to speed progress. At about the same time, Florida implemented a change in its A+ Plan for Education that focused on the educational gains of "low-performing" students. These two systems provide incentives for schools to concentrate differently on students even though they both ostensibly focus attention on similar sets of students – those most likely to be marginalized in public education. In this paper the authors study whether either of these accountability systems improved the academic outcomes of black, Hispanic and economically disadvantaged students in Florida. The authors find evidence that schools that are labeled as failing or near-failing in Florida's system tend to boost performance of students in these subgroups, while schools presented with incentives under NCLB to improve subgroup performance appear to be much less likely to do so. However, Hispanics appear to benefit from the NCLB sub-grouping requirements if they attend schools with low accountability pressure under Florida's grading system.

I. Introduction

Increasing the achievement of economically disadvantaged students as well as that of students of racial and ethnic minority groups is one of our nation's top education priorities. Indeed, according to the 2007 *National Assessment of Education Progress* 43% of 4th grade white students scored at or above the "proficient" level on the reading test; 51% scored at or above the proficiency level on the math test. This compares with only 14% and 17% of black and Hispanic students (respectively) scoring so highly in reading and 15% and 22% of black and Hispanic students scoring so highly in math.[1] Economically disadvantaged children, who themselves are disproportionately black or Hispanic, perform at similarly low levels. These disturbingly low levels of achievement are of national concern. Neal and Johnson (1996) and Tyler, Murnane and Willett (2000) document that student test scores are significantly related to subsequent educational attainment and labor market success. Closing such racial, ethnic and economic gaps in achievement – by raising the achievement of low-performing students – has been the focus of national educational policy for the past 40 years (most recently reflected in the *No Child Left Behind Act of 2001* (NCLB)) as well as countless state and local initiatives. The question is whether such accountability systems have, indeed, improved the performance of racial and ethnic minorities, or economically disadvantaged children, in the United States.

A growing literature examines the impact of accountability pressure on average student achievement with somewhat mixed results. For example, recent nationwide studies by Carnoy and Loeb (2002) and Hanushek and Raymond (2005) find significant improvement in student outcomes as a result of standards-based accountability, whereas the results from some specific

[1] Data are found at: http://nces.ed.gov/nationsreportcard/.

state systems have been less positive (see, e.g, Koretz and Barron (1998), Clark (2003) and Haney (2000, 2002)). To date, however, there is less evidence on which students are benefiting, and which may be losing, from these systems. Exceptions include Neal and Whitmore (2007) and Krieg (forthcoming) who suggest that proficiency-count-based systems[2] lead to concentration on the students in the middle of the distribution at the expense of both lower-achieving and higher-achieving students. Chakrabarti (2006) finds evidence that in Florida schools focused on students below the minimum thresholds for the school to be identified as "low performing." Further, Grissmer and Flanagan (1998) report that achievement gaps narrowed over the 1990s between advantaged and disadvantaged students in North Carolina and Texas, likely as a result of the accountability systems. Thus, while the literature is growing, we know relatively little about the impact of school accountability systems on subgroups of students, particularly the achievement of minority and disadvantaged students.

In this paper, we examine the impact of Florida's accountability system – the A+ Plan for Education – and NCLB on minority and disadvantaged student achievement. Florida's system enlists stigma (the grading of schools on an "A" to "F" scale), oversight (by the state of Florida), and competition to spur school improvement. Recent work by Figlio and Rouse (2006), Rouse et al (2007), Chiang (2007), Chakrabarti (2006), and West and Peterson (2006) have consistently found that student achievement significantly increased following a school's receipt of an "F" grade, presumably because of the increased accountability pressure. Whether the achievement of minority and disadvantaged students, per se, also increases when a school is awarded an "F" grade is unclear. One reason why the A+ Plan may contribute to the closing of the racial, ethnic, and economic achievement gaps is because such children are more likely to attend schools that

[2] As we discuss later, Florida's original system, implemented in 1999, is one such system.

receive failing or near-failing grades. In addition, because beginning in 2002 Florida's accountability system puts substantially greater weight on the performance of previously low-performing students than on that of other students, there may be an increased focus on such students generating improvements within schools.[3]

NCLB also requires that schools meet or exceed minimum proficiency levels in both math and reading. Plus, in order to make "adequate yearly progress" (AYP),[4] these same standards must also be met for every "subgroup" – where students are grouped by their race or ethnicity, English language proficiency, low-income and disability status. NCLB aims to raise the achievement of all students and to close racial, ethnic and economic gaps in achievement by exposing lagging achievement levels by certain populations. While fledgling, the empirical research on the impact of NCLB on student achievement is not particularly encouraging. For example, West and Peterson (2006) find that student achievement in Florida did not improve following a school's identification as failing to make AYP. They attribute the lack of improvement to the fact that a majority of schools in Florida were identified as failing to make AYP, diluting any stigma associated with the designation. More generally they argue that the incentives to improve and the consequences for "failing" are just too benign under NCLB. As a second example, Kane and Staiger (2002) estimate the impact of school accountability systems in Texas and California – that employ subgroup rules such as those in NCLB – on minority student achievement. They find no evidence that minority student achievement is greater in schools with identified subgroups compared to those where the minority student presence is not large enough

[3] Indeed, Rouse et al (2007) find that schools that received an F grade in 2002 were more likely to adopt policies that focused on low-performing students.

[4] A school is considered to be making "adequate yearly progress" (AYP) if all subgroups

to be identified.[5] They are careful to point out that such accountability systems may improve student achievement, overall, but that the subgroup rules do not have their intended impact.

To our knowledge, no research yet exists concerning whether the package of incentives and sanctions under NCLB, including explicit subgroup requirements, benefit minority or disadvantaged students to a greater or lesser degree than does the package of incentives and sanctions under the A+ Plan, including the general incentive to improve low-performing students regardless of subgroup. We exploit the fact that schools face differential accountability pressure under each of the two systems in an attempt to directly compare the performance effects of the two accountability systems. We find that schools subject to accountability pressure in Florida boosted the performance of racial minorities and economically disadvantaged children. On the other hand, the act of expecting schools to meet performance standards for specific subgroups à la NCLB did not generally lead to large improvements in the measured subgroups. However, students in identified subgroups may benefit from the NCLB subgrouping requirements if they attend schools with low accountability pressure under Florida's grading system. This last result is only seen, however, for Hispanic students but not for black students.

II. Two Paths to School Accountability

A. The Florida School Accountability Program

---

of students are achieving at the pre-defined proficient or advanced levels of achievement.

[5] As described more fully below, schools are only held accountable for the performance of students in subgroups if there are more than a specified minimum number of such students. Thus, schools with fewer than the specified minimum number of students in a subgroup are not held accountable for the performance of that subgroup per se, although the subgroup's members still count in measures of the overall performance of the school.

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 recommendations on how to improve, mandates that districts allocate certain resources and targeted funding for these schools, and priority for a program that provides reading coaches trained in scientifically-based reading research.[6] On the sanction side 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 – were eligible for school vouchers, called "Opportunity Scholarships." These vouchers allowed students to attend a different (higher rated) public school, or an eligible private school.[7] In addition, poor-performing schools were subject to additional scrutiny and oversight. All "D" and "F"-graded schools are subject to site visits and required to send regular progress reports to the state.

Between 1999 and summer 2001, schools were assessed primarily on the basis of aggregate test score *levels* and only in the grades with existing statewide curriculum-based

[6] See Rouse et al (2007) for more details on provisions of the A+ Plan.

[7] While in effect for nearly 7 years, the Opportunity Scholarship Program was declared unconstitutional by Florida's Supreme Court in January 2006. Other components of the A+ Plan, however, remain in effect.

assessments.[8] As can be seen in Figure 1, racial and ethnic minorities were disproportionately represented in poorly-graded schools. Starting in summer 2002, however, 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 2001-02 school year several things were known about the school grades that were to be assigned in summer 2002 (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 was not announced until the middle of the 2001-02 academic year, leaving schools with little time to adapt to its components. This relative "surprise" is a key component of our analytic strategy. As can be seen in Table 1, the distribution of school grades changed substantially from one system to the next, and Rouse et al. (2007) present evidence to suggest that these changes in the grade distribution are mainly due to changes in the system rather than to changes in school attributes or behaviors. That said, black and Hispanic students remained more heavily represented in poorly-graded schools, as shown in Figure 1.

This change in school accountability provided numerous incentives for schools. While the earlier system provided schools with the incentive to boost marginal students' performance and to potentially attempt to strategically alter the pool of students taking the test[9], the newer system evaluates schools on test score gains from one year to the next – and especially the gains of the

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

[9] Schools could alter the characteristics of students taking the tests through, for example, disciplinary actions (Figlio 2006) or reclassification of students (Figlio and Getzler 2006).

low-performing students. This emphasis on gains dramatically reduces the incentive for strategic behaviors, and encourages schools to care about the entire population of students, and particularly those students who are performing at low levels.

At the same time, the new system also takes the pressure off of schools that are performing at the highest levels. Whereas all schools faced accountability pressure under the old system, when the grading system was changed it also removed accountability pressure from a large number of schools. The reduction of accountability pressure for "A"-graded schools is evident in Table 2, which presents the likelihood that an "A" school would fall to a lower grade in the old versus new grading systems. The old grading system kept the pressure on all schools in part due to its uncertainty. Of schools that received a grade of "A" in 1999, 54 percent received a grade of "B" or below, and 25 percent received a grade of "C" or below, the next year. Fully 79 percent of "A" schools in 1999 had received a grade of "B" or below within two years of the "A" grade. In contrast, only 12 percent of "A" schools in 2002, the first year of the new system, would score a "B" or below the next year, and just one percent would score a "C" or below. For schools scoring 20 or more points above the "A" threshold in 2002, the percentage that would score a "B" or below the next year fell to 6 percent. It is reasonable to expect that schools facing much less accountability pressure might pay less attention to their minority and disadvantaged students, or for that matter, any other students that are low-performing. And even for those schools facing high accountability pressure, there is nothing to guarantee that schools will focus attention on their minority and disadvantaged students per se.

B. No Child Left Behind (NCLB)

Florida's school grading system and NCLB share numerous similarities. Both systems concentrate on the same grade levels in elementary school (the school level that we are considering in the present paper) – grades three and higher – and both systems focus attention on the same subject areas – mathematics and reading – using the same criterion-referenced state test, the Florida Comprehensive Assessment Test (FCAT). But they have some important differences as well: Florida's accountability system evaluates schools based on both the learning gains of students in the school and the proficiency rates of those students. NCLB, in contrast, focuses solely on the percentage of students in a school who are proficient. And while Florida's accountability system considers learning gains or test scores from black students and white students, for example, to be equivalent for the purposes of measuring school performance, NCLB explicitly requires that a school meet performance standards for each subgroup with a sufficient number of students.[10] The subgroups include students from economically disadvantaged families, those from major racial and ethnic groups, disabled students, and students with limited English proficiency. The performance goals for each subgroup increase over time, with the ultimate goal of 100 percent proficiency by the 2013-14 school year. Schools that meet these increasing performance goals for every subgroup in both reading and mathematics are said to be making "adequate yearly progress" (AYP).

Schools that fail to make adequate yearly progress are subjected to increasingly stringent sanctions, although many have questioned whether the sanctions have much bite. For example, if a school fails to make its AYP goals for two consecutive years, then the school is identified for school improvement and the district must allow students to transfer to another public school (in

[10] NCLB allows states to set their subgroup requirements for measurement. Florida requires that a school have 30 test-takers in any given subgroup for that subgroup to be counted

the district) that has not been identified as failing.[11] However, nationally only about 1 percent of eligible students actually exercise their choice option generating little or no competitive pressure for the "failing" schools to improve (Institute for Education Sciences (2006)).[12] If the school fails to make AYP for three consecutive years, in addition to the previous remedies, the district must allow parents to choose supplementary education services from providers with a "demonstrated record of effectiveness." However, Sunderman (2007) finds that student demand for such services leveled off or declined (even as more students have become eligible for such services) in districts after 2004. This lack of growth may be because parents are unaware of such services or because such services are not effective.[13] In either case, again it is not clear that facilitating student access to these private education providers generates serious competitive pressure for the schools.

It is clear that if a state introduces high performance standards, as Florida has, that a large fraction of schools will fail to make AYP. And indeed, this is precisely what happened: Three-quarters of Florida's schools did not make AYP in the first year of designation. The combination of this "blunt instrument" for grading schools, the identification of measured subgroups, and the

---

for school accountability purposes.

[11] In addition, the school must adopt a plan to improve performance in core academic subjects and the district must provide technical assistance to aid the school.

[12] Explanations for the anemic demand for choice include the fact that in districts with chronically failing schools, often there are few non-failing schools for students to attend and that the administration of the choice option is difficult for parents to understand and navigate. See Hannaway and Cohodes (2007) for a study of why take-up is so low in Miami-Dade County schools.

[13] As the ultimate sanction, if a school fails to make AYP for five consecutive years, then in addition to the previous remedies the district must restructure the school by reopening it as a charter school, replacing all relevant staff, entering into a contract with a private company to operate the school, or initiating a state takeover of the school. However, this is the first year that any school would have potentially been subject to this sanction such that there is little evidence

high rates of failing to make AYP leads to mixed incentives for schools. On the one hand, schools should face particular pressure to concentrate on every subgroup in the school, rather than just average performance, since the only way for schools to meet AYP expectations is if they boost performance in every subgroup. On the other hand, the high rates of failure to make AYP might cause schools to ignore the AYP designation altogether, especially if the general public does not pay attention to the ratings. Indeed, there exists very low concordance between Florida's school grades and the NCLB designations: In 2003, 55 percent of Florida's "A" schools and 87 percent of Florida's "B" schools failed to make AYP according to the federal standards. The "A" schools that failed to make AYP generally were more heterogeneous and had larger numbers of "countable" subgroups: Just over one-third of the "A" schools with 7 or 8 subgroups met the federal standards, while the federal pass rate for the "A" schools with fewer than four subgroups was nearly twice as high – almost 60 percent. Among high-performing schools, according to the state's designations, the more diverse a school was, the more likely it was to be punished under NCLB.

While this discordance between Florida's school grades and NCLB could lead parents and schools to discount the school ratings in one or both of these systems, it also indicates that NCLB could put accountability pressure on highly-rated schools that face little or no accountability pressure under Florida's grading system. Schools that have sufficiently high performance overall that they need not worry specifically about certain subgroups' performance under Florida's plan are forced to pay attention to these subgroups under NCLB, lest they be branded with a failing label according to the federal system. Therefore, NCLB could provide incentives for schools to boost the test performance of students in traditionally marginalized subgroups, especially when

---

on its effectives.

they are in schools that are otherwise high-performing. While the Florida system appears most likely to put performance pressure on the bottom-rated schools, NCLB is most likely to pressure the higher-rated schools to improve, at least along the dimension of the performance of racial, ethnic and economic subgroups.

Table 3 presents details on the percentage of schools that have sufficient numbers of students in a subgroup such that it would count for AYP determination, broken down by the school's grade. As can be seen, every "F"-graded school has a sufficiently large number of black and economically disadvantaged students for those subgroups to count for AYP purposes, but only 15 percent of elementary "F" schools have sufficient numbers of white students and 29 percent have sufficient numbers of Hispanic students for those subgroups to contribute individually to AYP. In contrast, among the "A"-graded elementary schools, 44 percent do not have a sufficiently large number of black students, 46 percent do not have sufficient Hispanic students, and 6 percent do not have sufficient white students for those subgroups to count for AYP purposes. And while the overwhelming majority of elementary schools have enough economically disadvantaged students for that subgroup to count for determining AYP, four percent of "A" schools do not. It is possible that NCLB could provide incentives for highly-rated schools to pay close attention to these subgroups even when the Florida grading system does not.

III. Data and Empirical Approach

A. Data

We rely on administrative data on individual elementary school students throughout the state (including all standardized test scores) from 1999-2000 through 2004-05. The data have

been longitudinally linked across years allowing us to follow students over time, as long as they do not leave the public school system or the state of Florida. (Those who leave and return do show up in the dataset and thus are still tracked over time.) From 2000-01 onward, all students in grades three and above took the criterion-referenced Florida Comprehensive Assessment Test (FCAT), while in 1999-2000, only fourth graders took the reading test and fifth graders took the mathematics test. In all of these years, all students in grades three and above also took the nationally norm-referenced Stanford 10 test.[14] In addition to test score results, these data also contain some individual-level characteristics such as the student's race, sex, eligibility for the National School Lunch Program (a measure of the student's socio-economic status since it is a means-tested program), and English-Language-Learner and disability status. In our present paper, we focus only on grades three through five, the traditional elementary school grades that are tested in Florida.

In each grade and year, around 200,000 students statewide take the FCAT. Few students are lost in the longitudinal analysis of these data: for instance, nearly 95 percent of students who took the fourth-grade FCAT are observed taking the FCAT the next year. Over the six year window, we observe 1,580,030 student-year observations for students eligible to receive free or reduced-price lunch (our measure of economic disadvantage), 711,159 black student-year observations, 640,580 Hispanic student-year observations, and 1,539,907 white student-year observations.

[14] To ease exposition, we focus exclusively on the FCAT rather than the norm-referenced test in this paper, since both accountability systems are designed to boost performance on the criterion-referenced test and success in both systems is measured against this performance. That said, the basic pattern of results presented herein is also observed with the norm-referenced tests.

B. Empirical Approach

To estimate the impact of Florida's A+ Plan and the federal NCLB on minority and disadvantaged student achievement levels, we estimate separate models for each of the three major racial and ethnic groups in Florida (blacks, Hispanics and whites) as well as for free and reduced-price lunch-eligible students. For each group of students, we estimate school fixed effects models in which the potential effects of school grades in 2002 and NCLB "turn on" beginning in the 2002-03 school year. Our specific estimation model is,

$$T_{ist} = \alpha_s + \beta_1 A_{s2002} POST_t + \beta_2 B_{s2002} POST_t + \beta_3 D_{s2002} POST_t + \beta_4 F_{s2002} POST_t + \gamma SUB_{s2002} POST_t + \tau_t + \varepsilon_{ist}$$

where T_{ist} represents the test score of student *i* in school *s* in year *t*;[15] A_{s2002}, B_{s2002}, D_{s2002}, and F_{s2002} are dummy variables indicating the school's accountability grade in 2002 (where a grade of "C" is the omitted category); SUB_{s2002} is a dummy variable indicating that the school has an identifiable subgroup for the subgroup of students in the regression (i.e., black, Hispanic, white, or free or reduced-price lunch eligible); $POST_t$ is a dummy variable indicating if the test score is from a year after the 2002 change in the grading formula (i.e., 2002-03, 2003-04, or 2004-05) – such that $A_{s2002}POST_t$ represents the interaction between A_{s2002} and $POST_t$, for example; α_s are school dummies; τ_t are year dummies; and ε_{ist} is assumed to be a normally distributed error term.[16] We also control for a vector of student characteristics (i.e., sex, disability status, eligibility for the National School Lunch Program, and grade in school).

[15] We standardize test scores in any given grade and year to have mean zero and a standard deviation of one. To reduce the potential for measurement error and to simplify discussion, we average the standardized reading score together with the standardized mathematics score to present a single test score.

[16] We cluster the standard errors at the school level.

One set of parameters of interest are the β coefficients as they represent the change in the given racial, ethnic or economic group's test scores following the formula change in the A+ Plan; we estimate four separate parameters, each compared with receipt of a "C" grade. Our other parameter of interest, γ, reflects the change in test scores associated with that subgroup now "counting" for that school's AYP status.[17] We also estimate models in which we treat the A-plus grades and the NCLB AYP designation effects in separate regressions.[18]

While schools were aware at the beginning of the 2002-03 school year that their AYP measures would depend on the scores of certain subgroups, it may be the case that they did not yet begin responding to this information until after the schools received their first AYP designations in 2003. Therefore, we also estimate alternative models in which we consider 2002-03 part of the pre-NCLB period such that $POST_t$ represents the 2003-04 and 2004-05 school years.

Finally, as mentioned above, the two accountability systems may have interactive effects. Perhaps low-ranked schools have aggregate performance that tends to be sufficiently far below the AYP standards that they are not at all motivated by NCLB to improve – or at least to improve the performance of specific measurable subgroups, but highly-rated schools that do not face much

[17] We do not control for lagged test scores because our fundamental interest is in understanding racial, ethnic and economic test score gaps, rather than test score changes per se, and because estimating effects on test score changes would require us to cut our sample to look only at students who are measured entirely in the pre-change period as compared with students whose lagged test scores are in the pre-change period but their ultimate test scores are in the post-change period. However, the estimated effects of the A-plus plan on test score levels presented herein are highly consistent with the overall estimated effects of the A-plus plan on test score gains presented in Rouse et al. (2007), suggesting that our choice of model specification is not influencing our findings.

[18] In addition, we control for the number of students in the relevant subgroup, so that the AYP subgroup effect can be interpreted as a regression-discontinuity effect. We note, however, that this modeling decision does not substantively influence the findings that we report. Results

pressure to improve by Florida's plan may still face pressure to avoid being labeled as failing under the federal system. In such a case, one would expect different types of schools to concentrate on measurable subgroups in different ways. To gauge the degree to which this is occurring, we estimate models such as:

$$T_{ist} = \alpha_s + \beta_1 A_{s2002} POST_t + \beta_2 B_{s2002} POST_t + \beta_3 D_{s2002} POST_t + \beta_4 F_{s2002} POST_t + \gamma SUB_{s2002} POST_t + \\ + \psi_1 A_{s2002} SUB_{s2002} POST_t + \psi_2 B_{s2002} SUB_{s2002} POST_t + \psi_3 D_{s2002} SUB_{s2002} POST_t + \\ \psi_4 F_{s2002} SUB_{s2002} POST_t + \tau_t + \varepsilon_{ist}$$

where the ψ parameters reflect the differential effect of school grades (relative to a grade of "C") when making AYP requires meeting performance targets for the specific subgroup in question.

IV. Results

A. Changes in Racial Gaps in Florida

Figure 2 presents the racial and ethnic test score gaps in elementary schools in Florida from 2000 to 2005 in both math and reading; these gaps represent the difference in test scores between minority students and white students.[19] The solid lines represent the "raw" gaps while the dashed lines represent the gaps after controlling for the student's sex, disability status, free lunch status, and grade in school. For the black-white and Hispanic-white gaps in achievement, the trends are remarkably consistent across the two tests: for both math and reading, there has been a narrowing between the achievement of black and white students over the 6 years, both with and without consideration of student characteristics. Examination of Figure 2 suggests that

---

turn out to be very similar if we exclude these controls.

[19] The test scores represent percentiles that have been standardized within Florida to have a standard deviation of 1. As a result, the gaps are in standard deviation units.

there is reason to suspect that the accountability pressure facing Florida's schools – either from Florida's new accountability system or from NCLB – just might be working to close racial and ethnic gaps in achievement.

Table 4 presents evidence on changes in black and Hispanic test scores, relative to those of white students, in Florida between 1999-2000 and 2004-05. As was seen in the figures, black and Hispanic students on average gained significantly at around the same time as the policy changes in 2002. While prior to the policy changes in 2002, blacks averaged one-half of a standard deviation lower test scores than did whites (after controlling for the handful of covariates available in the data) and Hispanics averaged one-third of a standard deviation lower test scores than did whites; following 2002 blacks and Hispanics each gained five percent of a standard deviation relative to whites. While the racial and ethnic test score gaps remain very large, they noticeably shrunk in a short period of time. The question is whether NCLB and/or changes in the school grading system in Florida are responsible for these changes.

B. Estimated Effects of Grading Changes and AYP Subgroup Rules

We next turn to the question of whether the two major policy changes in 2002 influenced the test scores of racial and ethnic minorities and economically disadvantaged students. As can be seen in Figures 3a and 3b, black, Hispanic and economically disadvantaged students appear to have gained ground in the years following 2002 in schools that were poorly graded – "D" or "F" – by the state. However, as seen in Figure 3c, no such gains occurred on average when the AYP subgrouping requirement was put into place. We estimate these changes more formally in Table 5a, where we model the estimated effects of the grading change in Florida, and in Table 5b where we study the impact of the AYP subgroup rules. In Table 5a, we observe consistent evidence that

minorities and economically disadvantaged students' test scores increased when their schools were given low grades, and especially grades of "F."[20] White students' grades also apparently increased, though not by as much, when the schools they attended received low grades, though the estimates are imprecisely estimated. White students apparently gained ground relative to minorities when their schools received grades of "A," a result consistent with the notion that highly-rated schools tended to focus less on historically disadvantaged student groups. In contrast, we find little evidence that schools with measurable black, economically disadvantaged, or white subgroups boosted the performance of the relevant subgroups, regardless of the timing of the estimated response, as shown in Table 5b. At the same time, schools with sufficient Hispanic students did appear to increase Hispanic student performance at the time that NCLB subgroup requirements would have encouraged them to do so.

In Table 6 we estimate a model that includes both sets of policies in the same specification. The same basic patterns emerge: Test scores increased the most in schools graded "F," with blacks benefiting most of all; test scores also increased substantially in "D"-graded schools; and subgrouping requirements apparently benefited Hispanic students, though the estimated effect of subgrouping requirements is half of the estimated effect of receiving a "D" grade and one-third of the estimated effect of receiving an "F" grade.[21] Similarly-sized estimates of the subgrouping requirement emerge for white students as well, but are not statistically distinct from zero.[22]

---

[20] The estimated effect of grade "F" receipt for Hispanics is less precisely estimated than the estimated effect for grade "D" effects, and is significant at only the 14 percent level.

[21] The estimated effect of the subgrouping requirement is not statistically distinct from the "D" grade effect at the ten percent level, but it is distinct from the "F" grade effect.

[22] We have estimated these same models using 2003-04 as the first post-NCLB year with

These basic findings suggest that while NCLB subgrouping requirements may have led to modest increases in Hispanic students' test scores, the school grading system apparently improved Hispanic students' scores (among students in those schools) by a larger amount. That said, since the majority of Hispanic students attend schools with countable Hispanic subgroups but only 7 percent of Hispanic students attend schools graded "D" or "F," it is likely that the NCLB subgroup requirement was more influential than the change in the school grading system in improving Hispanics' overall scores post-2002.[23] There is no such ambiguity for black students or for economically disadvantaged students: We could find no evidence that the subgroup requirements benefited these students, while there exists strong evidence that black and disadvantaged students attending low-ranked schools improved their scores following the change in the grading system.

As described above, highly-rated schools may have the incentive to pay less attention to minorities and economically disadvantaged students once they face lower accountability pressure. NCLB subgrouping requirements might provide new accountability incentives for these schools. To gauge the degree to which this might be the case, we interact school grades with measured subgroup indicators. The results of this exercise are presented in Table 7.[24] In this table, we further subdivide "A"-graded schools into "marginal A" schools – those scoring fewer than 20 points above the "A" grade threshold – and "safe A" schools – those scoring 20 or more points

---

very comparable results.

[23] It is impossible to gauge how influential the accountability system writ large has been on test scores. For example, students in disadvantaged subgroups in "C" schools, the comparison group in the grading system, gained two to three percent of a standard deviation in the period following the change in the grading system. Therefore, estimated effects of the change in grading can probably be thought of as a lower bound estimate.

[24] We focus only on black and Hispanic students in this exercise because nearly all Florida elementary schools are held accountable for the performance of economically disadvantaged

above the threshold for "A" receipt in 2002. We find that the differences between schools with measurable subgroups and those without measurable subgroups never statistically significant for black students.[25] However, we observe that "safe A" schools explicitly held accountable for Hispanic student performance have Hispanic students who fare significantly better than do "safe A" schools without this requirement for making AYP. The other differences for Hispanic students are not statistically significant. This evidence suggests that the estimated benefits of the AYP subgroup requirement for Hispanic students is concentrated in those schools that would potentially be expected to have reduced accountability pressure under Florida's grading system. We will pursue this issue in greater depth in future versions of this paper.

V. Conclusion

This paper presents new evidence on the effectiveness of two forms of school accountability systems in raising the academic performance, at least in terms of standardized test scores, of historically low-performing student subgroups. We find strong evidence that labeling schools as failing or near-failing leads to improved performance of black, Hispanic and economically disadvantaged students. While explicitly holding schools accountable for black

---

students.

[25] Since every "F" school also is required to meet a black subgroup requirement, it is impossible to know with certainty how much of the "F" grade result is due to the subgroup requirement and how much is due to the "F" grade receipt. That said, while they considered only overall effects of accountability rather than those for specific subgroups, Rouse et al. (2007) demonstrate a sharp regression discontinuity in "F" grade receipt in a model in which all but four comparison schools had a measurable black subgroup. We have replicated this finding, restricting our analysis to the set of comparison schools with measurable black subgroups, and the estimated effect of receiving a grade of "F" persists unchanged, and is quite similar in magnitude to the results presented herein. Therefore, there is strong reason to believe that the "F" effect shown here is due to the grading system and not to the black subgrouping requirement.

students' scores does not apparently lead to gains for black students, we do find that explicitly holding schools accountable for the performance of Hispanic students appears to improve Hispanic students' scores. Upon further analysis, we find that these gains are concentrated in what we call "safe A" schools – those that had arguably been released from accountability pressure following the change in the grading system in 2002.

The results of this research indicate that NCLB's requirement that schools meet a series of subgroup hurdles combined with rather ineffective sanctions is not likely to lead to large improvements in the performance of historically disadvantaged students – at least in the case of states such as Florida with high standards for proficiency and heterogeneous schools with many subgroups and therefore many chances to fail to meet AYP standards. In a state of the world in which the typical school fails to meet the standards with few penalties, it stands to reason that many schools would be unresponsive to the accountability pressure put forth by NCLB. In contrast, holding schools responsible for student learning gains, and especially the learning gains of low-performing students, combined with real stigma, increased oversight, competition appears to lead to substantial gains in the progress of minority and economically disadvantaged students. Note that these results may also suggest that it is possible to improve the achievement of disadvantaged and minority students without singling out their performance for accountability purposes, at least when focusing on low-performing schools.

That said, a grading system such as Florida's provides little incentive to boost student performance in schools that are "destined" to receive a very high grade. NCLB subgrouping requirements may put pressure on these schools to focus on at least some students historically left behind when the standard grading system does not, though the evidence on this point is currently weak. These findings, if they hold up to closer scrutiny, may suggest that the ideal school

accountability system would still put accountability pressure on high-performing schools to improve the performance of minorities and economically disadvantaged students.

References

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): 533-575.

Boozer, Michael A., Alan B. Krueger, and Shari Wolkon. "Race and School Quality Since Brown Versus the Board of Education," in *Brookings Papers on Economic Activity: Microeconomics* Martin N. Bailey and Clifford Winston (eds.) (Washington, DC: The Brookings Institution, 1992), pp. 269-338.

Carnoy, Martin and Susanna Loeb (2002). "Does External Accountability Affect Student Outcomes? A Cross State Analysis." *Education Evaluation and Policy Analysis,* 24(4): 305-331.

Chakrabarti, Rajashri (2006) "Vouchers, Public School Response and the Role of Incentives: Evidence from Florida." Working Paper, Federal Reserve Bank of New York.

Chiang, Hanley. "How Accountability Pressure on Failing Schools Affects Student Achievement," Harvard University mimeo (October 2007).

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

Ferguson, Ronald F. "Teachers' Perceptions and Expectations and the Black-White Test Score Gap," *Urban Education*, 38, no. 4 (July 2003), pp. 460-507.

Figlio, David. (2006) "Testing, Crime and Punishment." *Journal of Public Economics* 90(4): 837-851.

Figlio, David and Lawrence Getzler (2006). "Accountability, Ability and Disability: Gaming the System?" in Improving School Accountability: Check-Ups or Choice, Advances in Applied Microeconomics, T. Gronberg and D. Jansen, eds., 14 (Amsterdam: Elsevier Science).

Figlio, David and Cecilia Rouse (2006). "Do Accountability and Voucher Threats Improve Low-Performing Schools?" *Journal of Public Economics* 90(1-2): 239-255.

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.

Grissmer, David and Ann Flanagan. "Exploring Rapid Achievement Gains in North Carolina and Texas," Paper prepared for the National Education Goals Panel, November 1998.

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

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

Hannaway, Jane and Sarah Cohodes. "Miami-Dade County: Trouble Even in Choice Paradise," in *The No Child Left Behind Remedies: Safe? Sensible? Effective?* Frederick M. Hess and Chester E. Finn Jr. (editors) (Washington D.C.: American Enterprise Institute Press, 2007).

Hanushek. Eric et al. (1994) *Making Schools Work: Improving Performance and Controlling Costs*. Washington, DC: The Brookings Institution.

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

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

Kane, Thomas J. and Douglas O. Staiger. "Racial Subgroup Rules in School Accountability Systems." Dartmouth College mimeo, September 2002.

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

Krieg, John (forthcoming). "Are Students Left Behind? The Distributional Effects of the No Child Left Behind Act." *Education Finance and Policy*.

National Assessment of Title I: Interim Report (Washington D.C.: Institute of Education Sciences, National Center For Education Evaluation and Regional Assistance, 2006).

Neal, Derek and Diane Whitmore Schanzenbach (2007) "Left Behind by Design: Proficiency Counts and Test-Based Accountability." Working paper, National Bureau of Economic Research, August.

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

Tyler, John H., Richard J. Murnane, and John B. Willett. "Do Cognitive Skills of School Dropouts Matter in the Labor Market?" *The Journal of Human Resources*, 35, no. 4 (Autumn 2000): 748-754.

Sunderman, Gail L. "Supplemental Educational Services under NCLB: Charting Implementation" The Civil Rights Project, UCLA, Policy Brief (October 2007). (http://www.civilrightsproject.ucla.edu/research/esea/SES_Policy_Brief.pdf)

Sunderman, Gail L, James S. Kim, Gary Orfield. (editors) *NCLB Meets School Realities: Lessons from the Field* (Thousand Oaks, CA: Corwin Press, 2005)

West, Martin and Paul Peterson (2006). "The Efficacy of Choice Threats Within School Accountability Systems: Results from Legislatively Induced Experiments," *The Economic Journal* 116: C46-C62.

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	1235	1203
B	299	255	399	549	565	515
C	1180	1115	1074	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	2297	2289	2330	2501	2501	2490
	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	1555	1543	1610	1670	1670	1667

Source: Authors' calculations from state data.

Table 2: Probabilities that an A-graded School Earned a Lower Grade in Subsequent Years: Old and New Systems

	Probability of scoring "B" or below			Probability of scoring "C" or below		
	Year after	Two years after	In either year	Year after	Two years after	In either year
Schools with "A" grade in summer 1999	0.54	0.46	0.79	0.25	0.17	0.38
Schools with "A" grade in summer 2002	0.12	0.14	0.22	0.01	0.02	0.03
Schools 20+ points above "A" threshold in summer 2002	0.06	0.06	0.11	0.01	0.01	0.02
Schools 50+ points above "A" threshold in summer 2002	0.05	0.02	0.06	0.01	0.01	0.02

Source: Authors' calculations from state data.

Table 3: Percentage of Schools with Measurable Subgroups for AYP Calculation in 2002-03

School Grade	Subgroup			
	Black	Hispanic	White	Economically disadvantaged
	All Schools			
A	65%	60%	96%	97%
B	77%	64%	92%	99%
C	86%	63%	84%	100%
D	93%	59%	52%	100%
F	98%	41%	24%	100%
Overall	76%	61%	86%	98%
	Elementary Schools			
A	56%	54%	94%	96%
B	71%	59%	89%	99%
C	81%	54%	76%	100%
D	93%	49%	42%	100%
F	100%	29%	15%	100%
Overall	70%	54%	82%	98%

Source: Authors' calculations from state data.

Table 4: Changes in Racial and Ethnic Test Score Gaps in Florida, 2000-05

Model specification	Coefficient estimate			
	Black	Black x Post 2002	Hispanic	Hispanic x Post 2002
No additional covariates	-0.768 (0.010)	0.053 (0.006)	-0.501 (0.013)	0.039 (0.007)
Controlling for sex, free/reduced price lunch, and disability status	-0.477 (0.007)	0.051 (0.005)	-0.338 (0.008)	0.052 (0.006)
Allowing covariates to have different coefficients before vs. after the policy change	-0.477 (0.007)	0.051 (0.005)	-0.338 (0.008)	0.052 (0.006)

Notes: Each row represents a separate regression in which the dependent variable is the average standardized FCAT reading and mathematics score; note that white non-Hispanic is the omitted race/ethnicity category. Standard errors adjusted for clustering at the school level are in parentheses beneath point estimates.

Table 5a: Estimated Effects of Changes in FL Grading System on Student Achievement

School grade in 2002	Subgroup			
	Black	Hispanic	White	Economically disadvantaged
A	-0.001 (0.012)	-0.001 (0.013)	0.014 (0.007)	-0.001 (0.009)
B	-0.006 (0.013)	0.009 (0.015)	0.001 (0.008)	-0.002 (0.010)
D	0.034 (0.018)	0.051 (0.023)	0.036 (0.023)	0.044 (0.015)
F	0.109 (0.036)	0.069 (0.047)	0.026 (0.077)	0.101 (0.032)

Table 5b: Estimated Effects of NCLB Subgroup Requirements on Student Achievement

	Subgroup			
	Black	Hispanic	White	Economically disadvantaged
Subgroup counted for AYP (effect starting in 2002-03)	0.010 (0.015)	0.026 (0.016)	0.013 (0.020)	0.014 (0.048)
Subgroup counted for AYP (effect starting in 2003-04)	0.001 (0.015)	0.029 (0.016)	0.005 (0.021)	0.020 (0.054)

Notes: Each column represents a separate regression in which the dependent variable is the average standardized FCAT reading and mathematics score. Standard errors adjusted for clustering at the school level are in parentheses beneath point estimates. Regressions also control for year dummies, school fixed effects, and student characteristics.

Table 6: Estimated Effects of Changes in Grading System or NCLB Subgroup Requirements

School grade in 2002	Subgroup			
	Black	Hispanic	White	Economically disadvantaged
A	-0.010 (0.012)	-0.006 (0.013)	0.007 (0.008)	0.006 (0.009)
B	-0.011 (0.013)	-0.002 (0.014)	-0.003 (0.009)	-0.003 (0.010)
D	0.036 (0.018)	0.051 (0.022)	0.040 (0.022)	0.041 (0.015)
F	0.116 (0.036)	0.082 (0.048)	0.053 (0.080)	0.112 (0.033)
Subgroup counted for AYP (effect starting in 2002-03)	0.001 (0.015)	0.025 (0.016)	0.020 (0.020)	-0.016 (0.048)

Notes: Dependent variable is the average standardized FCAT reading and mathematics score. Standard errors adjusted for clustering at the school level are in parentheses beneath point estimates. Regressions also control for year dummies, school fixed effects, and student characteristics.

Table 7: Estimated Combined Effects of School Grades And NCLB Subgroup Requirements For Schools with Different Grades

	Subgroup					
	Black students			Hispanic students		
School grade in 2002	Schools with measurable subgroup	Schools without measurable subgroup	p-value of difference	Schools with measurable subgroup	Schools without measurable subgroup	p-value of difference
"Safe" A – 430 points or higher	-0.026 (0.027)	-0.042 (0.029)	0.447	0.036 (0.027)	-0.007 (0.029)	0.061
"Marginal" A – 410-429 points	-0.043 (0.028)	-0.034 (0.036)	0.773	0.034 (0.028)	0.040 (0.039)	0.875
B	-0.034 (0.027)	-0.038 (0.034)	0.902	0.038 (0.027)	0.025 (0.035)	0.646
D	0.012 (0.030)	-0.002 (0.093)	0.877	0.091 (0.033)	0.083 (0.046)	0.864
F	0.092 (0.044)	n/a	n/a	0.115 (0.057)	0.156 (0.087)	0.680

Notes: Dependent variable is the average standardized FCAT reading and mathematics score. Standard errors adjusted for clustering at the school level are in parentheses beneath point estimates. Regressions also control for year dummies, school fixed effects, and student characteristics.

Figure 1: Distribution of Students across School Grade, by Race and Year

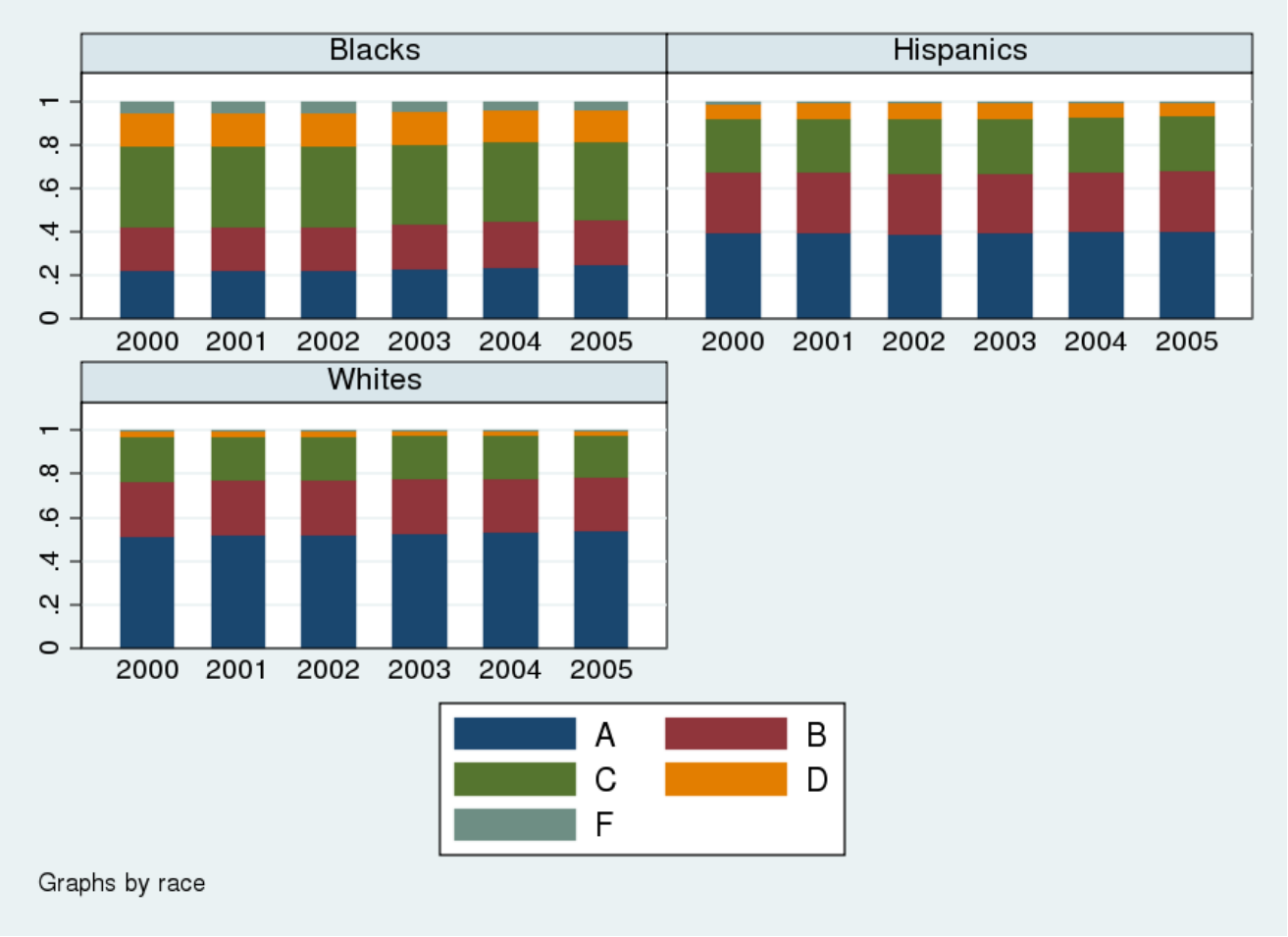


Figure 2: Raw and Residual Test Score Gaps Between Black/Hispanic and White Students

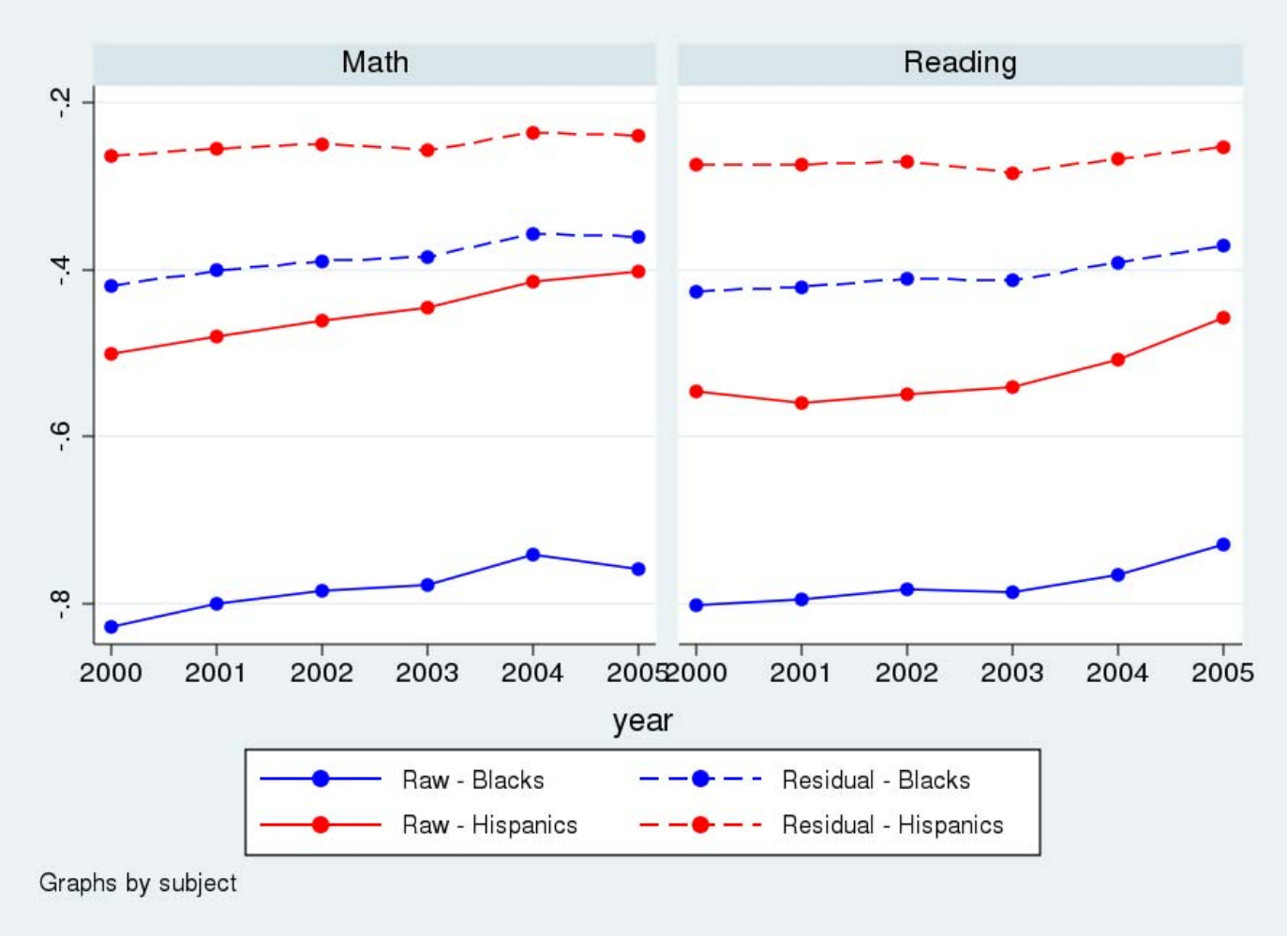


Figure 3a: Over-Time Changes in Average Test Scores of Different Racial/Ethnic Groups, by 2002 School Grade

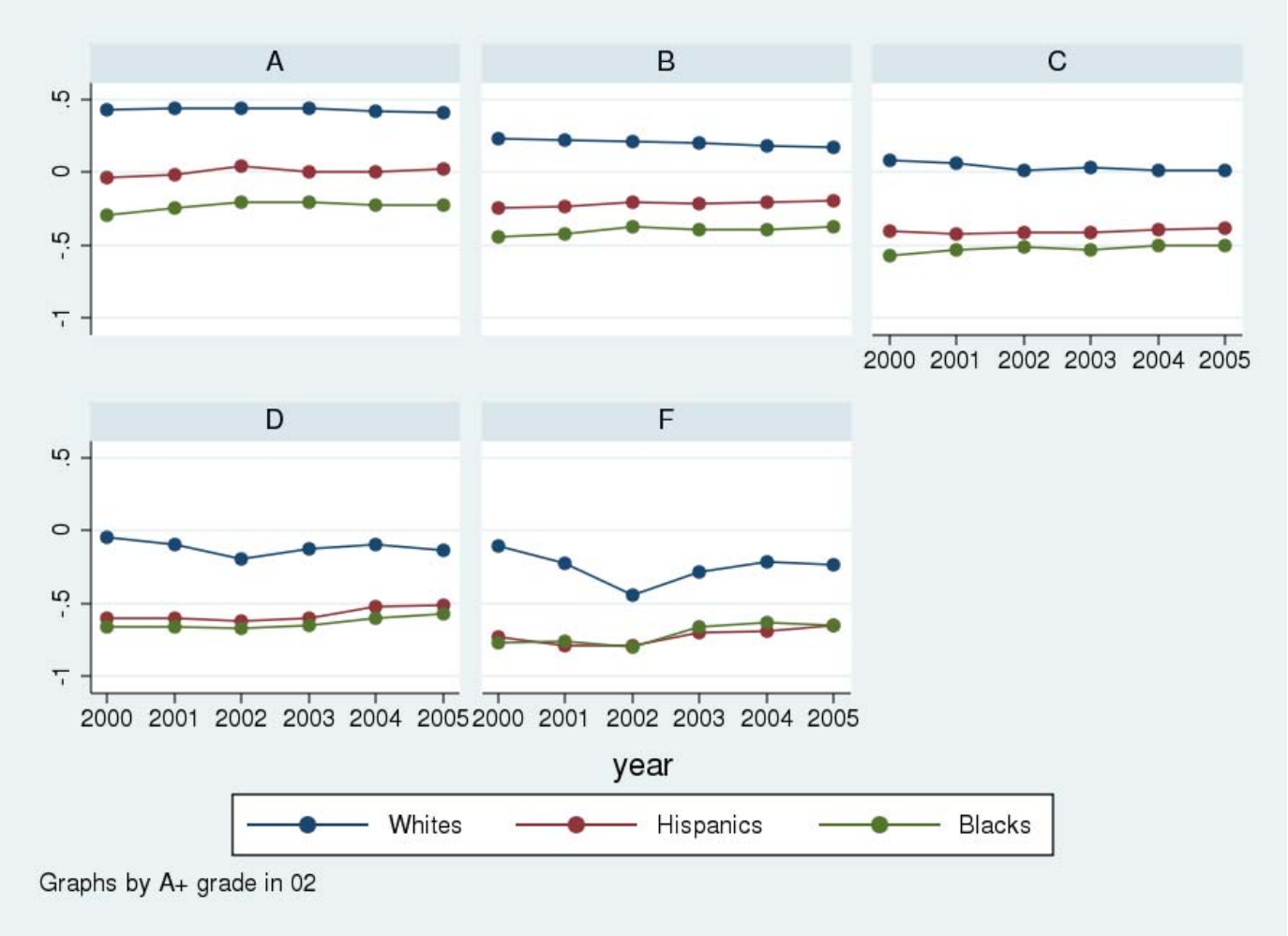


Figure 3b: Over-Time Changes in Average Test Scores of Subsidized and Non-subsidized Lunch Students, by 2002 School Grade

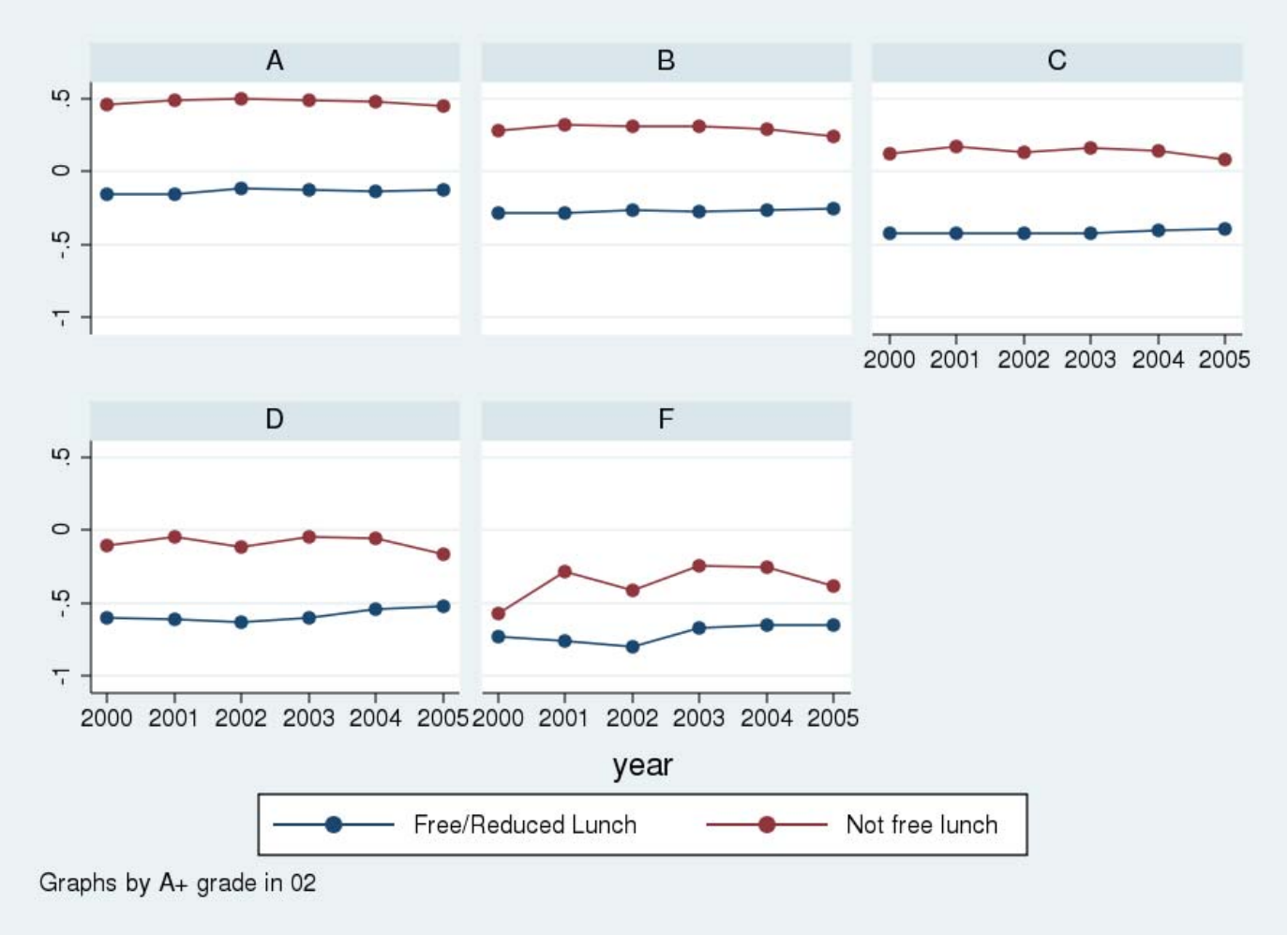
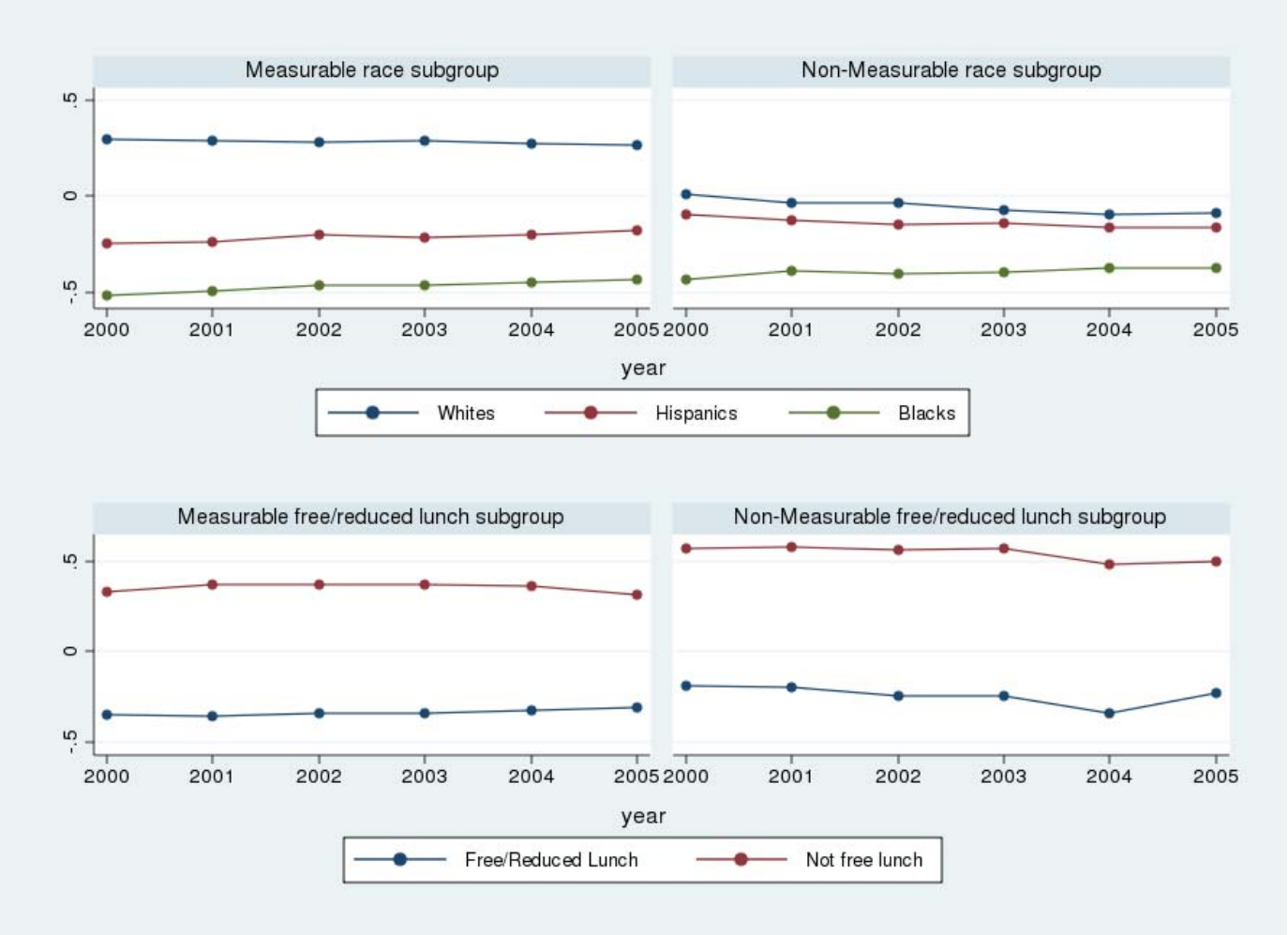


Figure 3c: Over-Time Changes in Average Test Scores of Different Groups, For Schools with and without the Relevant Measurable Subgroups




CALDER