LEFT BEHIND BY DESIGN: PROFICIENCY COUNTS AND TEST-BASED ACCOUNTABILITY

Derek Neal and Diane Whitmore Schanzenbach*

Abstract—We show that within the Chicago Public Schools, both the introduction of NCLB in 2002 and the introduction of similar district-level reforms in 1996 generated noteworthy increases in reading and math scores among students in the middle of the achievement distribution but not among the least academically advantaged students. The stringency of proficiency requirements varied among the programs implemented for different grades in different years, and our results suggest that changes in proficiency requirements induce teachers to shift more attention to students who are near the current proficiency standard.

> *We were told to cross off the kids who would never pass. We were told to cross off the kids who, if we handed them the test tomorrow, they would pass. And then the kids who were left over, those were the kids we were supposed to focus on.*
>
> —Middle school staff member (de Vise, 2007)†

I. Introduction

FOR more than a decade, test-based accountability systems have been a key element of many education reform proposals at the state and district levels, and the No Child Left Behind Act (NCLB) of 2001 created a federal mandate for test-based accountability in every state. A key feature of NCLB is the requirement that each state adopt an accountability system built, in large part, on standardized testing in reading and math for students in grades 3 through 8. The law seeks to hold schools accountable for student performance by mandating that schools make the results of these standardized assessments available to parents and report not only aggregate results but also results specific to particular demographic groups, such as groups defined by race or special education status. These reports must convey the fractions of students in particular schools and demographic groups within schools who have achieved proficiency in a particular subject for their grade level. NCLB spells out a set of sanctions that schools should expect to face if they persistently report proficiency levels below the targets set by their state for each calendar year.

In this paper, we use data from the Chicago Public Schools (CPS) to examine how a specific aspect of the implementation of NCLB affects the distribution of measured changes in achievement among students. The implementation of NCLB in most states and the design of many state and local accountability systems tie rewards and sanctions to the number of students in certain groups scoring above given proficiency thresholds. We use the introduction of two separate accountability systems in CPS, a district-wide system implemented in 1996 and the introduction of NCLB in 2002, to investigate the impacts of proficiency-count accountability systems on the distribution of student performance.

In all our analyses, we focus on test score outcomes among students in a given grade. We compare students who took a specific high-stakes exam under a new accountability system with students who took the same exam under low stakes the year before the accountability system was implemented. Furthermore, because we restrict our comparisons to students who took exams either right before or right after the implementation of an accountability system, we can make these comparisons holding constant student performance on a similar low-stakes exam in an earlier grade. Thus, we are able to measure changes in test scores associated with the accountability system at different points in the distribution of prior achievement.

Holmstrom and Milgrom (1991) warn that when workers perform complex jobs involving many tasks, pay-for-performance schemes based on objective measures of output often create incentives for workers to shift effort among the various tasks they perform in ways that improve their own performance rating but hinder the overall mission of the organization. Holmstrom and Milgrom cite "teaching to the test" in response to test-based accountability systems as an obvious example of this phenomenon, and much of the existing empirical literature on test-based accountability focuses on whether the test score increases that commonly follow the introduction of such systems represent actual increases in subject mastery. The literature explores many ways that schools may seek to inflate their assessment scores without actually increasing their students' subject mastery. Schools may coach students for assessments, manipulate the population

Received for publication February 12, 2008. Revision accepted for publication September 18, 2008.

† Quote from an anonymous middle school staff member in "Rockville School's Efforts Raise Questions of Test-Prep Ethics" by Daniel de Vise, *Washington Post*, March 4, 2007.

* Neal: University of Chicago and NBER; Schanzenbach: University of Chicago and NBER.

We thank Elaine Allensworth, John Q. Easton, and Todd Rosenkranz of the Consortium on Chicago School Research for their assistance in using the data. We thank Amy Nowell of Chicago Public Schools (CPS). We thank participants in the University of California, Berkeley, Labor Economics Workshop, the Federal Reserve Bank of Chicago's Labor Economics seminar, the Harris School's Public Policy and Economics Workshop, and the joint meeting of the Institute for Research on Poverty's Summer Research Workshop and the Chicago Workshop on Black-White Inequality. We thank Fernando Alvarez, Gadi Barlevy, Kelly Bedard, Julie Berry Cullen, Jennifer Jennings, Brian Jacob, John Kennan, Roger Myerson, Kalina Michalska, Phil Reny, and Balazs Szentes for useful comments and discussions, and Chloe Hutchinson, Garrett Hagemann, Richard Olson, and Andy Zuppann for helpful research assistance. We owe special thanks to Phil Hansen for being so generous with his time and his knowledge of accountability within CPS. D.N. thanks the Searle Freedom Trust for generous research support as well as Lindy and Michael Keiser for support through the University of Chicago's Committee on Education. Both thank the Population Research Center of NORC and the University of Chicago for research support.

of students tested, or even alter students' answer sheets between assessment and grading.[1]

We depart from most of the existing literature by examining a different multitasking concern. Instead of focusing on how the use of standardized assessments shapes what teachers teach and what types of coaching they give their students, we examine how the rules that accountability systems use to turn student test scores into performance rankings for schools determine how teachers allocate their efforts among different students. We show that the use of proficiency counts as performance measures provides strong incentives for schools to focus on students who are near the proficiency standard but weak incentives to devote extra attention to students who are either already proficient or have little chance of becoming proficient in the near term.

Even in a world with perfect assessments that cannot be manipulated by schools in any way, the details of how one maps students' test scores into a performance rating for their school dictate how teachers allocate their attention among students of different baseline achievement levels. Because part of the impetus for NCLB and related reforms is the belief that some groups of academically disadvantaged students have historically received poor service from their public schools, we believe that our results speak to a design issue that is of first-order importance.[2]

We provide results that characterize the distribution of test score changes among fifth graders in Chicago following the introduction of NCLB in 2002, and we present similar results for fifth graders tested in Chicago in 1998 following the introduction of a school accountability system that was similar to NCLB on many dimensions. The results for both sets of fifth graders follow a strikingly consistent pattern. Students at the bottom of the distribution of measured third-grade achievement score the same or lower following these reforms than one would have expected given the prereform relationships between third- and fifth-grade scores, but students in the middle of the distribution score significantly higher than expected. Further, there is at best mixed evidence of gains among students in the top decile.

We also present results for sixth graders tested in 1998. These students were affected directly by both the school-level accountability system instituted within CPS and a separate set of test score cutoffs used to determine summer school placement and retention decisions. Chicago's effort to end social promotion linked summer school attendance and retention decisions to score cutoffs that were much lower than the proficiency cutoffs used to determine school-level performance. Thus, sixth graders who had little chance of contributing to their school's overall proficiency rating faced strong incentives to work harder in school. The results for these sixth graders follow the same general pattern observed in the fifth-grade results. However, the estimated gains among sixth graders tend to be larger at each decile, and our estimated treatment effects for the least able sixth graders are never negative. We conclude that NCLB provides relatively weak incentives to devote extra attention to students who have no realistic chance of becoming proficient in the near term or students who are already proficient.[3]

The distributional consequences of the Illinois implementation of NCLB are complex. Hanushek and Raymond (2004) argue based on National Assessment of Educational Progress (NAEP) data and differences over time and among states in the stakes associated with state-level accountability systems that test-based accountability reduces racial achievement gaps, and our results are not inconsistent with this conclusion. The CPS contain relatively few white students, and average test scores did increase following both NCLB and the CPS reforms of 1996. Thus, although we do not have comparable data from other school districts in Illinois, our results certainly admit the possibility that NCLB narrowed the achievement gaps between whites and minorities in the state. However, the group of students within CPS who were likely not helped and may have been harmed by NCLB is sizable and predominantly black and Hispanic.

Several studies on the use of proficiency counts in accountability systems other than NCLB provide results that are consistent with ours. Reback (2008) uses data from Texas during the 1990s to measure how schools allocated efforts in response to a statewide accountability system. He finds that achievement gains are larger among students whose gains are likely to make the greatest marginal contribution to their school's overall proficiency rating. Burgess et al. (2005) use data from England to show that achievement gains are lower among less able students if they attend schools in which a large fraction of the student body are marginal students with respect to an important score threshold in the English accountability system. On the other hand, Springer (2008) analyzes data from the testing program that Idaho instituted following the introduction of NCLB and argues that he does not see evidence that the use of profi-

[1] See Carnoy and Loeb (2002), Grissmer and Flanagan (1998), Hanushek and Raymond (2004), Jacob (2005), and Koretz (2002). These studies address the concern that teaching to the test artificially inflates test scores following the introduction of high-stakes testing. See Cullen and Reback (2006) for an assessment of strategic efforts among Texas schools to improve reported scores by manipulating which students are exempt from testing. Jacob and Levitt (2003) provide evidence that some teachers or principals in Chicago actually changed student answers after high-stakes assessments in the 1990s.

[2] Lazear (2006) notes that the scope of assessments may also influence the distribution of gains among students. Those who find learning difficult may not be affected if assessments are too broad because they and their teachers may find it too costly for them to reach proficiency.

[3] NCLB does contain a provision that requires that all students be proficient by 2014. However, this provision of the law does not constitute a credible threat. NCLB contains a reauthorization requirement for 2007 and has still not been reauthorized. Goals that push the limits of credulity and are not required by NCLB until 2014 should play a small role in shaping teachers' and principals' expectations concerning how the law will be enforced today.

ciency counts in NCLB led to increased teacher effort among only one particular group of students.

All of these papers differ from ours methodologically because none of the authors had access to data on achievement growth prior to the introduction of accountability. We have test score data on all Chicago students starting in the early 1990s, and the tests used for NCLB purposes in Illinois and those used for the district's accountability program in the late 1990s were administered in years prior to the introduction of these accountability systems. Thus, ours is the only study in this literature with access to control groups that took the assessments used in accountability systems as low-stakes exams prior to the introduction of accountability.[4]

Several studies of particular schools also find results consistent with those we present below. Gillborn and Youdell (2000) coined the term *educational triage* to describe their findings from case studies of English schools. They document how these schools targeted specific groups of students for special instruction in order to maximize the number of students who performed above certain thresholds in the English system. More recently, Booher-Jennings (2005) and White and Rosenbaum (2007) present evidence from case studies of two schools serving economically disadvantaged students in Texas and Chicago, respectively. Both studies provide clear evidence that teachers and administrators made conscious and deliberate decisions to shift resources away from low-performing students and toward students who had more realistic chances of exceeding key threshold scores.

In the next section, we present a model of teacher effort within schools. Then we turn to the details of the 1996 and 2002 reforms and their implementation in Chicago before turning to our empirical results. After presenting our results, we discuss the challenges that policymakers face if they wish to replace NCLB's reliance on proficiency counts with a system of measuring progress that will value the achievement gains of all students. Currently a number of states have been granted waivers that allow them to assess school performance using more continuous measures of student outcomes than simple proficiency counts. We analyze the likely effects of these alternative schemes using variants of the same model of teacher effort that we describe in the next section. Our model clearly illustrates that these waivers make it easier to design accountability systems that do not build in direct incentives to leave some children behind, but we argue that tough design issues remain unresolved. In our conclusion, we discuss the extent to which our results from Chicago speak to the likely effects of NCLB in other large cities.

[4] The Idaho and Texas data provide no information about how various schools performed in the absence of NCLB. Springer (2008) notes that his results may reflect "customary school behavior irrespective of NCLB's threat of sanctions."

II. Keeping Score Using Proficiency Counts

Consider a school that is part of a test-based accountability system. Two policies shape the actions of teachers and principals. To begin, the central administration, in cooperation with parents, provides enough monitoring to make sure the school provides some baseline level of instruction to all students. Because our empirical work measures changes in performance that follow the introduction of accountability within groups of students who are similar with respect to prior achievement levels, it is not essential for our purposes that baseline instruction be identical for all students, but this assumption does allow us to easily describe both our model and our empirical results in terms of the changes induced by accountability systems. We do not address the socially optimal amount of effort per teacher or the socially optimal allocation of effort among students. We take as given for now that teachers and the principal enjoy rents given their pay and the baseline allocation of effort per student. Thus, we view the introduction of test-based accountability as an attempt to extract more effort from teachers, and we examine how this attempt to increase overall teacher effort also changes the distribution of teacher effort among students of different abilities.

Given the monitoring system that guarantees baseline effort, also consider a testing system that labels each student as either passing or failing. Further, assume that the principal and teachers incur costs that are a function of the number of their students who fail. These costs may take many forms depending on the details of the accountability system.[5] The key point is that NCLB keeps score, and the earlier Chicago accountability system kept score, based on the number of students whose test scores exceed certain thresholds. Thus, we model our hypothetical accountability system as a penalty function that imposes costs on teachers and principals when students do not reach a proficiency standard, and we assume that these costs are strictly convex in the number of students who fail.[6]

Our school can improve individual test scores by providing extra instruction beyond the minimum effort level that the district can enforce through its monitoring technology.

[5] Under NCLB, schools must report publicly how many of their students are proficient, and they face serious sanctions if their proficiency rates remain below statewide targets. In Chicago, the district adopted a system in 1996 that measured school-level performance based on the number of students exceeding national norms on specific exams. In addition, Chicago schools and students faced additional pressures related to a separate set of lower thresholds (on the same tests) that determined whether students in grades 3, 6, and 8 were required to attend summer school and possibly repeat their grade.

[6] NCLB also includes provisions concerning the fraction of students who are proficient within certain demographic groups defined by race, family income, and disability status. Incorporating these subgroup provisions in our model would complicate our analyses but not change our results. The high cost of bringing low-achieving students up to proficiency would still imply that schools could optimally allocate no extra instruction to their low-achieving students. Further, these provisions are less important for NCLB implementation in Chicago than many other school districts because schools in Chicago are highly segregated by race and income.

We ignore any agency problem between principals and teachers and model the school as a unitary decision-making unit.

Because the baseline level of instruction for all students is not a choice variable for the school, the school's problem is to minimize the total cost incurred by the allocation of extra instruction among its students and the penalties associated with student failures. Suppose that there are N students in a school and each student has ability

$$\alpha_i,\ i = 1, 2, \ldots, N.$$

Further, assume that for any individual i, her score on the accountability test is

$$t_i = e_i + \alpha_i + \varepsilon_i$$

where e_i is extra instruction received by student i and ε_i is the measurement error on i's test drawn from $F(\varepsilon)$, which has a unimodal density $f(\varepsilon)$.

The cutoff score for passing is $\bar{t}$. We assume that N is large, and we approximate the school's objective function by treating the expected number of students who fail in each school as the actual number of failures in each school. Thus, the school's problem is as follows:

$$\min_{e_i} \Psi[\sum_{i=1}^{N} F(\bar{t} - e_i - \alpha_i)] + \sum_{i=1}^{N} c(e_i) \qquad (1)$$

$$\text{s.t. } e_i \geq 0 \ \ \forall\, i = 1, 2, \ldots, N.$$

Here, $\Psi[\cdot]$ is a penalty function that describes the sanctions suffered by a school of size N under the accountability system, and $c'(e) > 0$, $c''(e) > 0$, $\forall e \geq 0$. The penalty function is strictly increasing and convex in the number of students who are not proficient. The first-order conditions that define optimal effort require

$$\Psi'[\,\cdot\,] f(\bar{t} - e_i^* - \alpha_i) \leq c'(e_i^*) \ \ \forall\, i = 1, 2, \ldots, N.$$

The precise shapes of the penalty function, the cost function, the distribution of ability types, and the distribution of measurement errors interact to determine the exact pattern of optimal investments. However, we know that in any setting that involves convex cost and penalty functions as well as a unimodal density for the measurement error, the optimal investment pattern will exhibit two properties. First, it is easy to show the following: $\bar{t} - \alpha_i < \bar{t} - \alpha_j \Leftrightarrow \bar{t} - \alpha_i - e_i^* < \bar{t} - \alpha_j - e_j^*\ \forall i, j$. This means that optimal investments never cause one type to pass another in terms of effective skill. The responses of schools to this type of accountability system may narrow the achievement gaps between various skill groups, but these responses will not eliminate or reverse any of these gaps. We are not surprised that our empirical results firmly support this prediction,[7] but we do believe that it is important to recognize that the type of accountability system described here, which is intended to capture the key elements of NCLB, should not be expected to fully eliminate achievement gaps between any two groups of students.

Second, while it is easy to generate examples such that schools devote no extra effort to students below some critical ability level or above some critical ability level, appendix B demonstrates that solutions do not exist that involve a school's allocating no extra effort to a given student but applying positive extra effort to other students who are both more and less able. If the solution to the school's problem involves positive extra effort for some students and no extra effort for others, the students who do not receive extra attention will be either more or less skilled than those who do.

The focus of our empirical work is the claim that accountability systems built around proficiency counts provide incentives for schools to provide extra help to students in the middle of the ability distribution while providing few incentives for these schools to direct extra attention to students who are either far below proficiency or already proficient. We think that the absence of strong incentives to help students who are achieving at the lowest levels is especially noteworthy because this feature of the NCLB design is at odds with the stated goals of the legislation, and the implication that the least able may be left behind by design is a quite robust feature of our model. Although we have assumed that the marginal product of instruction is independent of student ability, we could make the more common assumption that ability and instruction are complements in the production of knowledge. In this scenario, the relative cost of raising scores among less able students increases, and it remains straightforward to construct scenarios in which students below a given ability level receive no extra attention even though more able students do benefit from the accountability system.

It is worth noting that under this type of accountability system, the choice of $\bar{t}$ determines the distribution of achievement gains. Consider an increase in the standard for proficiency $\bar{t}$. It is easy to show that this increase in the standard can only decrease and never increase the number of high-ability students who receive no extra instruction. Thus, higher standards can only benefit and never harm the most able students. However, a higher standard may actually increase the number of low-ability students that a given school ignores by increasing the number who have little or no chance of being proficient in the near term.[8] Although NCLB encourages each state to set challenging proficiency

[7] We define ability groups based on baseline achievement in previous grades, and our data appendix shows that average math and reading scores always increase from one ability group to the next for both our treatment and control cohorts.

[8] A higher standard does not necessarily generate this result. A higher standard also raises the baseline failure rate and, because the penalty

standards, states that set high standards may direct teacher effort away from disadvantaged children.

Also note that one can easily construct a more general model that embeds our analyses of effort allocation within schools as one component in a model of the labor market for teachers and principals. Here, differences among schools in the indirect utilities associated with the solutions to the effort allocation problems faced by various schools will drive the sorting of teachers and principals among schools. Assuming the function $\Psi(\cdot)$ is the same for all schools of the same size, schools with more able students provide a superior working environment for principals and teachers because academically disadvantaged students raise the cost of meeting any specific passing rate given a common proficiency standard. If the distribution of initial student ability is worse in school A than school B, teachers and principals in school A must work harder than those in school B to achieve the same standing under the accountability system, and this should adversely affect the relative supply of teachers who want to teach in school A.

There has been little empirical work to date on how accountability systems affect teacher labor markets, but Clotfelter et al. (2004) examine changes in teacher retention rates in North Carolina following the introduction of a statewide accountability system in 1996 that raised the relative cost of teaching in schools with large populations of disadvantaged students. They document significant declines in retention rates among schools with many academically disadvantaged students, and their results are difficult to square with the hypothesis that the additional departures from these schools were driven primarily by an increase in the departure of incompetent teachers.

III. High-Stakes Testing in Chicago

We use data in the years surrounding the introduction of two separate accountability systems in CPS. The first, implemented in 1996, linked school-level probation status to the number of students who achieved a given level of proficiency in reading on the Iowa Test of Basic Skills (ITBS). It also linked grade retention decisions concerning individual students in "promotion gate" grades to the achievement of specific proficiency levels in reading and math. The second system is the 2002 implementation of NCLB testing in Illinois, which initially covered student performance in grades 3, 5, and 8 on the Illinois State Achievement Test (ISAT).

During 1996, a new administration within the CPS introduced a number of reforms, and these reforms attached serious consequences to standardized test results.[9] In the fall of 1996, CPS introduced a school accountability system. Among elementary schools, probation status was determined primarily by the fraction of students who earned reading scores equal to or greater than the national norm for their grade. Schools on probation were forced to create and implement school improvement plans, and these schools knew that they faced the threat of reconstitution if their students' scores did not improve. Although math scores were not a major factor in determining probation status, schools also faced pressure to improve math scores. As part of the reform efforts, CPS chose to publicly report proficiency rates in math and reading at the school level. Principals and teachers knew that the reading and math performance of their students would be reported in local newspapers, and these school report cards measured school performance using the number of students who performed at or above national norms in reading or math. With regard to sanctions and public reports, proficiency counts were the key metric of school performance in the CPS system.

In addition, other score thresholds in reading and math played a large role in the reform. In March 1996, before the school accountability system was introduced, CPS announced a plan to end social promotion. The new elementary school promotion policy required students in third, sixth, and eighth grades to score above specific thresholds in math and reading or attend summer school. These cutoff scores were far below the national norms that CPS would later use to calculate proficiency rates for schools, but they were clearly relevant hurdles for students in the bottom half of the CPS achievement distribution. Even median students likely faced more than a 20% risk of summer school if they exerted no extra effort. Students who attended summer school were tested again at the end of summer and retained if they still had not reached the target score levels for their grade.

This policy was announced in late March 1996. CPS exempted third- and sixth-grade students from the policy until spring of 1997, but the new policy did link eighth-grade summer school and retention decisions to the 1996 spring tests results. Since the promotion policy was announced only weeks before testing began, we believe that the eighth-grade exams in spring 1996 do not reflect all the impacts of the reform, but the eighth-grade exams are affected by the March announcement.[10] Thus, we restrict our attention here to the fifth- and sixth-grade results.

The retention policies in the CPS reforms are interesting from our perspective because CPS also built these policies around cutoff scores and because retentions forced students and their families to deal with a summer school program that they did not choose. Thus, retentions represented a source of potential frustration for parents and another source of performance pressure linked to proficiency counts. Further, the lower cutoff scores for summer school put many students at risk of summer school while still giving almost

function is convex, raises the gain associated with moving any single student up to the proficiency standard.

[9] See Bryk (2003) and Jacob (2003) for more on the history of recent reform efforts in CPS.

[10] In related work, we have discovered that the school-level correlation between ITBS score and IGAP (Illinois Goals Assessment Program) scores dropped notably for eighth graders in 1996.

all students a real chance to avoid it. This was not the case with regard to the proficiency levels used to determine school-level performance under the 1996 reforms, and it was not the case with regard to the ISAT proficiency cutoffs under NCLB in 2002. Thus, the results for sixth-grade students allow us to see what the distribution of achievement gains looks like when more students have a realistic chance of meeting an important threshold score.

The 1996 CPS reforms adopted the ITBS as the primary performance assessment in reading and math. Different forms of the test were given in different years, but in our analyses of ITBS data, we concentrate only on years when Form L was given. These years, 1994, 1996, and 1998, are the only ones surrounding the 1996 reform that permit a comparison of prereform and postreform cohorts using a common form of the ITBS. Our analyses seek to measure changes in scores relative to prereform baselines at different points in the distribution of prior achievement. If we use years other than the Form L years, our results will reflect not only any real differences in the effects of the reform at various ability levels but also any differences among ability levels in the accuracy of the psychometric methods used to place scores from different forms on a common scale. While it is not easy to equate scores among forms in a manner that is correct on average, the task of equating scores in a manner that is accurate at each point in the distribution of ability is even more demanding.

In the 1998–99 school year, the Illinois State Board of Education (ISBE) introduced a new exam to measure the performance of students relative to the state learning standards and administered the test statewide, but only in grades 3, 5, and 8. For many reasons, CPS viewed the ISAT as a collection of relatively low-stakes exams during the springs of 1999, 2000, and 2001.[11] However, in fall 2001 with the passage of NCLB looming, the ISBE placed hundreds of schools in Illinois on a watch list based on their 1999 through 2001 scores on ISAT and also declared that the 2002 ISAT exams would be high-stakes exams.

By the time President Bush signed NCLB in early January 2002, it had become crystal clear that the 2002 ISAT would be the NCLB exam for Illinois. Further, the state announced in February that for the purpose of calculating how long each school had been failing under NCLB, 1999 would be designated as the baseline year and school status in the year 2000 would retroactively count as the first year of accountability. This meant that many schools in Chicago expected to start to face sanctions immediately if their proficiency counts on the 2002 spring ISAT exams did not improve significantly. Thus, in one year, the ISAT went from a relatively low-stakes state assessment to a decidedly high-stakes exam.

Like the 1996 CPS reforms, NCLB employs proficiency counts as the key metric of school performance. States are required to institute a statewide annual standardized test in grades 3 through 8, subject to parameters set by the U.S. Department of Education. States set their own proficiency standards as well as a schedule of target levels for the percentage of proficient students at the school level. If the fraction of proficient students in a school is above the goal, the school is said to have met the standard for Adequate Yearly Progress (AYP).[12] Under some circumstances, if a school does not have enough proficient students in the current year but does have a substantially higher fraction than in previous years, the school may be considered to have met the AYP standard under what is called the "Safe Harbor Provision." A school that persistently fails to meet the AYP requirement will face increasing sanctions. These include mandatory offering of public school choice and extra services for current students, and at some point, the school may face reconstitution.

We are not able to conduct our analyses of ISAT scores using a sample restricted to students who took the exact same form of the exam. ISBE typically administered ISAT using two forms simultaneously. These forms shared a large number of common items both within and across years, and thus the assessment program was designed in a manner that facilitated ISBE's use of an item response theory model to place all scores on a common scale from 120 to 200. We cannot control for any form effects in our ISAT analyses because the CPS data that we use do not allow us to determine which form a given student took in a given year. Nonetheless, an independent audit of the ISAT did conclude that ISAT scores are comparable over time and among forms of the exam.[13]

IV. Changes in Scores

All the figures presented in this section follow a common format. They display differences between mean test scores in a specific grade following the introduction of high-stakes testing and mean predicted scores based on data from the period prior to high-stakes testing. We create our estimation samples using selection rules that take the following form: we include persons who were enrolled in CPS in year t and

[11] The ISAT was not a "no-stakes" exam in 1999–2001. ISAT performance played a small role in the CPS rules for school accountability over this time, and the state monitored ISAT performance as well. Nonetheless, according to Phil Hansen, Chicago's former chief accountability officer, CPS began participating in ISAT under the understanding that the results would not be part of any "high-stakes accountability plan." In late fall 1999, the state made several announcements that signaled a change in this position, and CPS protested. Then, in January 2000, ISBE moderated its stance and informed CPS that it would appoint a task force to recommend a "comprehensive school designation system" for state-level accountability and a set of guidelines that would exempt schools with low ISAT scores from being placed on the state's Academic Early Warning List if they "show evidence of continued improvement." Thus, in the springs of 1999, 2000, and 2001, CPS took the ISAT with the expectation that the results would not have significant direct consequences in terms of the state accountability system.

[12] In addition, the fraction of students passing in each subgroup above a minimum size must meet the standard. For example, NCLB defines subgroups by race, socioeconomic status, and special education category.

[13] Wick (2003) provides a technical audit of the ISAT.

year $t + 2$ in grades n and $n + 2$ respectively, and we restrict our samples to students who were tested in math and reading in both years.[14] Appendix A provides a detailed description of how we construct our samples and the characteristics of our treatment and control samples. Relative to our control cohorts, we observe slightly higher rates of follow-up testing for the cohort affected by NCLB and slightly lower rates of following testing for the cohorts that experienced the earlier CPS reforms. However, in both cases, our prereform and postreform cohorts match well on baseline characteristics within our estimation samples, which we define by achievement decile, grade, and reform year.

With regard to our analyses of the CPS accountability system, the two-year intervals reflect the fact that 1994, 1996, and 1998 are years around the CPS reform that involve assessment using the same form of the ITBS. We present results for fifth and sixth graders because these are the cohorts tested in 1998 that did not face any promotion hurdles under the CPS reforms in 1996 or 1997. The two-year interval is also necessary in our analyses of the 2002 implementation of NCLB. ISBE administered the ISAT in only third, fifth, and eighth grades. We cannot analyze eighth-grade scores in the pre-NCLB period given controls for fifth-grade achievement because the ISAT was first administered in 1999, but we can use the third-grade scores from 1999 and the fifth-grade scores from 2001 to estimate the pre-NCLB relationship between ISAT scores in fifth and third grades.[15]

In all our analyses, we compare outcomes in a specific grade for two different cohorts of students. Both cohorts took tests in two grades, and both cohorts took their tests in the lower grade under low stakes. However, the latter cohort took exams in the higher grade under high stakes. For our ISAT results, these stakes reflect the Illinois 2002 implementation of NCLB. For our ITBS results, these stakes reflect the 1996 introduction of CPS's accountability system. Our goal is to examine how test scores in the higher grade change following the introduction of an accountability system based on proficiency counts controlling for achievement in the lower grade, and we are particularly interested in the possibility that the effects of accountability may differ among various levels of prior student achievement in the lower grade.

For the purpose of describing our estimation procedure, we refer to the cohorts tested in both grades under low stakes as the prereform cohorts and the cohorts tested under high stakes in the higher grade as the postreform cohorts. Our estimation procedure is as follows:

- We begin by using the prereform cohort to estimate the first principal component of math and reading scores in the baseline grade. This principal component serves as a baseline achievement index.
- We use the coefficient estimates from this principal component analysis and the lower-grade math and reading scores from the postreform cohort to construct indices of baseline achievement for students in the postreform cohort as well. These indices tell us where the postreform students would have been in the distribution of baseline achievement for the prereform cohort.
- We divide the pre- and postreform samples into ten cells based on the deciles of the prereform distribution of baseline achievement.
- Given these cells, we run twenty separate regressions. For each of our ten samples of prereform students, we run two regressions of the form

$$y_{igk} = c + \beta_1 y_{i(g-2)math} + \beta_2 y_{i(g-2)read} + \beta_3 (y_{i(g-2)math} * y_{i(g-2)read}) + u_{igk},$$

  where y_{igk} is the score of student i in grade g on the assessment in subject k.
- Based on these regression results, we form predicted scores, $\hat{y}_{igk}$, for each person in the postreform cohort and then form the differences between these predicted values, $\hat{y}_{igk}$, and the actual grade g scores in math and reading for the postreform cohort.
- Finally, we calculate the average of these differences in math and reading for each of our ten samples of students in the postreform cohort and plot these averages.[16]

A. *NCLB Results*

Figures 1A and 1B present our estimates of the changes in fifth-grade reading and math scores associated with the 2002 implementation of NCLB in Illinois. For students whose third-grade scores place them in the bottom two deciles of the 1999 achievement distribution, there is no evidence that NCLB led to higher ISAT scores in fifth grade. Three of the four estimated treatment effects for these deciles are negative. The only statistically significant estimated effect implies that fifth graders in 2002, whose third-grade scores placed them in the bottom decile of the

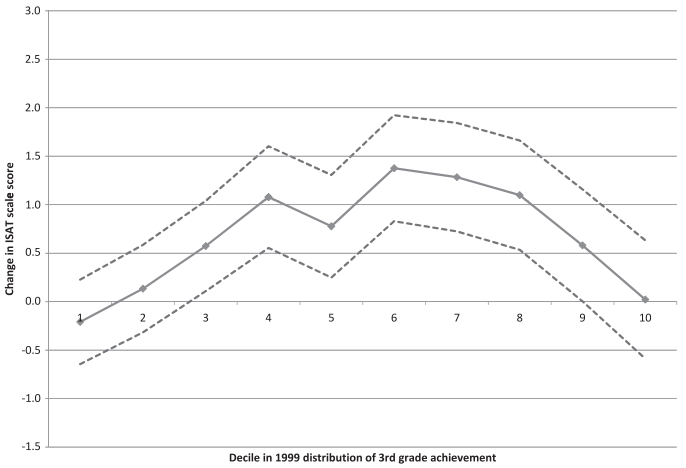
[14] We use the last year a student was in third grade as the third-grade year. We obtain similar results if we use test scores from the first year of third grade.

[15] In an earlier version of this paper, we also presented comparisons between the 1999–2001 cohort and the 2001–2003 cohort. However, we subsequently learned that the interval between the 2001 third-grade test and the 2003 fifth-grade test was at least nine weeks shorter than the intervals for the cohorts that we deal with here. While the patterns in these results are quite similar to those presented in figures 1A and 1B, we cannot rule out the possibility that the difference in time between assessments as well as other differences in test administration for the 2001–2003 samples affect those results.

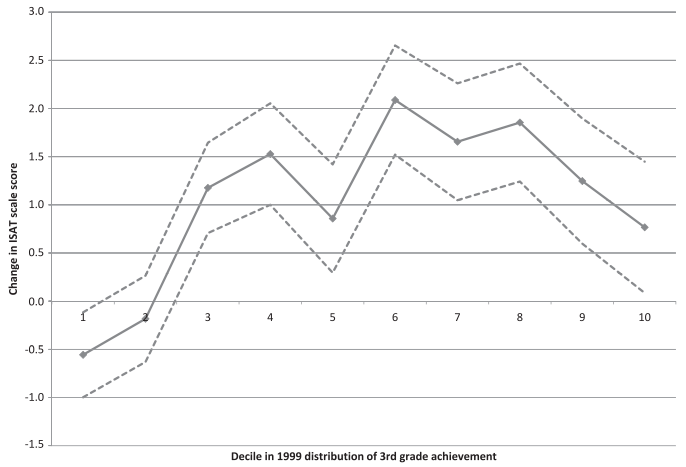
[16] The bands in the figures are 95% confidence intervals. We calculate these intervals accounting for the fact that we must estimate what the expected score for each student would have been in the absence of NCLB. We obtain the adjustments to the variances of our estimates of mean cell differences by taking the sample average of the elements of the matrix $(Z\hat{\Omega}Z')$ where N is the number of fifth-grade observations in 2002, Z is the $N \times 4$ matrix of third-grade score variables used to produce predicted scores, and $\hat{\Omega}$ is the estimated variance covariance matrix from the regression of 2001 fifth-grade math or reading scores on these third-grade variables from 1999.

FIGURE 1.—FIFTH-GRADE RESULTS FOR 2002

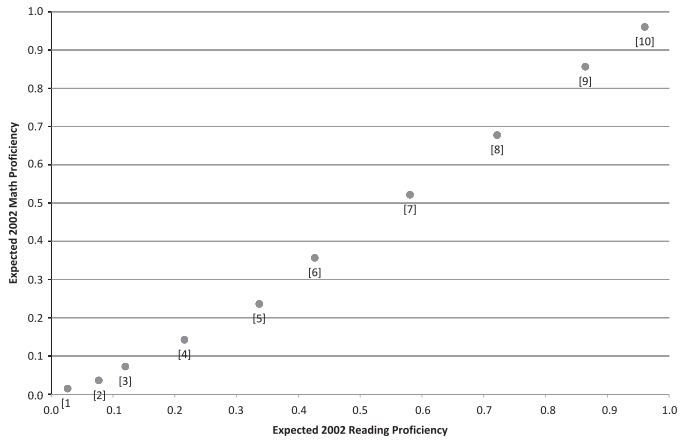
A. CHANGE IN FIFTH-GRADE READING SCORES, 2002 VERSUS 2001



B. CHANGE IN FIFTH-GRADE MATH SCORES, 2002 VERSUS 2001



C. EXPECTED 2002 PROFICIENCY IN FIFTH-GRADE BY DECILE OF THE THIRD-GRADE ACHIEVEMENT DISTRIBUTION, 1999



1999 third-grade achievement distribution, scored just over one-half point lower in math than expected given the observed relationship between third-grade scores in 1999 and fifth-grade scores in 2001. Because the ISAT scale is designed to generate a standard deviation of 15 for all scores, this estimated effect represents a decline of roughly 0.04 standard deviations. In contrast, deciles 3 through 9 enjoy higher-than-expected ISAT scores in both math and reading. We observe the largest score gains in math and reading in the sixth decile, where fifth graders in 2002 scored just under 0.1 standard deviations higher in reading and more than 0.13 standard deviations higher in math than comparable fifth graders scored in 2001.

Figure 1C presents the expected proficiency rates in math and reading for each of the deciles included in figures 1A and 1B.[17] These are the rates expected given the third-grade performance of students who were in fifth grade in 2002 and the relationship between third- and fifth-grade performance for the 2001 cohort of fifth graders. For example, the figure tells us that in the absence of NCLB, the fifth graders in 2002 who fell in the fifth decile of our baseline achievement distribution would have faced just over a 20% chance of reaching the proficiency standard for math and just under a 35% chance of reaching the reading standard.

In light of figure 1C, we are not surprised that we did not find an increase in ISAT scores in 2002 among students in the bottom two deciles. The Illinois proficiency standards are lofty goals for these students, and they face less than a 10% chance of reaching either standard. The fact that we do find significant positive effects for students in the third decile suggests that students may receive some benefit from these types of reforms even if they have at best modest hopes of reaching the threshold for proficiency. It is important to note that the model outlined in section II can accommodate this result. We assume that the cost function associated with investment in any particular student is convex. If small investments in students are rather inexpensive at the margin, schools may find it optimal to make such investments, even in students who are notably below the current proficiency standard.[18] On the other hand, our results for the third-decile students may reflect spillover effects that are not present in our model. In any event, figures 1A and 1C demonstrate that students with the lowest levels of prior achievement did not appear to achieve higher ISAT scores following NCLB, and among these students, the Illinois proficiency standards represented almost unattainable goals. Taken as a whole, these results support our contention that NCLB is not designed to leave no child behind.

B. Interpretation and Robustness of the NCLB Results

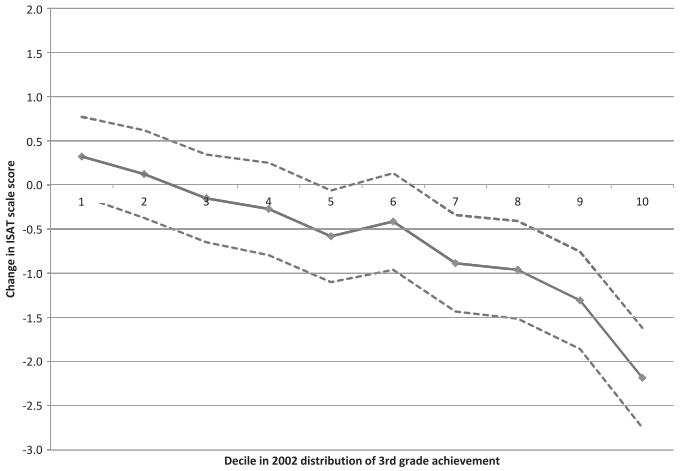
Several issues regarding the interpretation of our results deserve further attention. First, figures 1A and 1B present estimated changes in the scores on specific assessments. We can state clearly that the ISBE implementation of NCLB worked better, in terms of raising ISAT scores, for some

[17] These expected proficiency rates are predicted values based on the estimated coefficients from a logit model of fifth-grade proficiency in 2001 given third-grade math and reading scores in 1999.

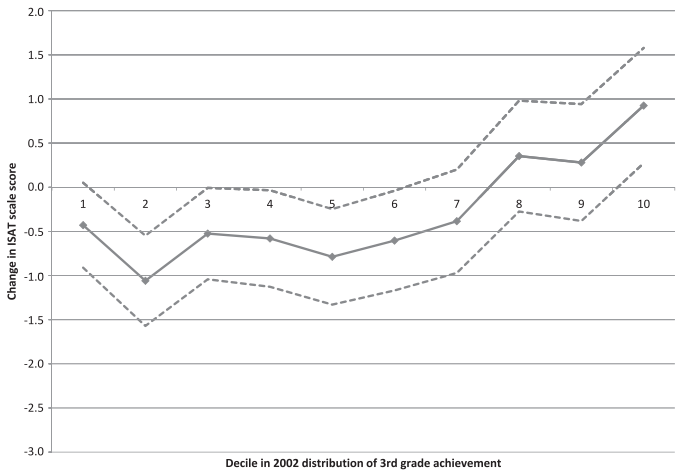
[18] Examine our first-order condition above. The value of $c'(e)$ for $e = 0$ will play a large role in determining how many students receive extra attention as a consequence of the accountability program.

FIGURE 2.—PLACEBO TESTS BASED ON 2005 RESULTS

A. CHANGE IN FIFTH-GRADE READING SCORES, 2005 VERSUS 2004



B. CHANGE IN FIFTH-GRADE MATH SCORES, 2005 VERSUS 2004



students than others and that it may have been counterproductive among the least able students in CPS; it is worth noting that this claim does not rest on a particular choice of scaling for the ISAT scores. We find no evidence of positive effects among students in the bottom two deciles but clear evidence of significant increases in ISAT scores among students in deciles 3 through 9. If all the estimated effects were the same sign, we might worry that any comparisons among cells concerning the magnitude of estimated effects could be sensitive to our choice of scale for reporting test scores, but our main emphasis here is a qualitative, not a quantitative, claim. Scores are higher than expected for students who are in the middle of the baseline achievement distribution and the same or lower than expected for those at the bottom of this distribution. Although NCLB raised average ISAT scores in Chicago, the implementation of NCLB in Chicago did not help, and may have hurt, the children who were likely the furthest behind when they began school. Our model suggests that this outcome should not be a surprise, but it is also not consistent with the stated purpose of NCLB.

We would like to conduct placebo experiments using ISAT data from the years before 2002 in order to rule out the possibility that we are simply picking up preexisting differences among ability levels in the trends of third- to fifth-grade changes in test scores among CPS students. However, this is not possible because only three years of ISAT data exist prior to 2002, and we need four years of data to measure differences in third- to fifth-grade achievement trajectories between two cohorts of students. Nonetheless, we can construct comparisons in reading and math using two cohorts tested under the same policy regime. The 2005 and 2004 cohorts of fifth graders were tested in both fifth and third grades under NCLB. Thus, we construct figures describing changes in fifth-grade scores between 2005 and 2004 in order to examine changes in scores between two cohorts tested under similar policy regimes. Figures 2A and 2B do not offer even a hint of the clear pattern that is observed in figures 1A and 1B. We see sizable losses in reading and some noteworthy gains in math among the top deciles, but there is no common pattern for math and reading results, and there is no evidence of important gains in the middle of the distribution relative to the lower deciles.

We do not know why there are some statistically significant deviations from 0 in these figures. In any pair of years, especially during the early years of a new policy regime, there may be differences in test administration or curricular priorities that create such differences.[19] Our main point is that these figures describe differences between two cohorts that experienced broadly similar accountability environments, and these differences in no way fit the pattern observed in figures 1A and 1B.

Table 1 contains the results from two different robustness checks. We perform these checks not only on our 2002 analyses but also on the 1998 analyses presented in the next section. In the first exercise, we add controls for race, gender, and free-lunch status to the regression models that we use to form predicted final grade scores. The second exercise involves adding school fixed effects to these models. The second exercise is not quite as straightforward as the first. Because there are over 400 elementary schools in Chicago and roughly 2,000 fifth-grade students per year in each of our baseline achievement deciles, many schools are represented in a given baseline achievement decile in 2002 but not in 2001. Thus, we cannot estimate separate regressions for each of our deciles and simply add school fixed effects without losing a significant number of observations.[20] Nonetheless, we can estimate the relationships between fifth- and third-grade scores for the 2001–1999 cohort

[19] NCLB is designed to be more demanding over time, and the target proficiency rate did increase modestly in Illinois between 2004 and 2005.

[20] We used four groups: deciles 1 and 2, 3 through 5, 6 through 8, and 9 and 10. We employed a richer polynomial in third-grade achievement scores to compensate for the use of four broader regression cells instead of ten. We still calculate average treatment effects for each decile to facilitate comparisons with our other results.

TABLE 1.—ROBUSTNESS CHECKS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Math						Reading					
	Main Effect	SE	With Demographic Controls	SE	With School Fixed Effects	SE	Main Effect	SE	With Demographic Controls	SE	With School Fixed Effects	SE
	A: 5th Grade ISAT (2002 vs. 2001), Grade-Equivalent Units											
Decile in 1999 3rd-grade achievement distribution												
1	−0.56	(0.23)	−0.46	(0.22)	−0.78	(0.23)	−0.21	(0.22)	−0.11	(0.22)	−0.37	(0.22)
2	−0.18	(0.23)	0.01	(0.22)	0.01	(0.22)	0.13	(0.23)	0.21	(0.23)	0.28	(0.22)
3	1.18	(0.24)	1.14	(0.23)	1.12	(0.24)	0.57	(0.24)	0.55	(0.24)	0.52	(0.24)
4	1.53	(0.27)	1.53	(0.26)	1.41	(0.26)	1.08	(0.27)	1.05	(0.26)	1.06	(0.25)
5	0.86	(0.29)	0.74	(0.28)	1.10	(0.27)	0.78	(0.27)	0.69	(0.27)	0.94	(0.27)
6	2.09	(0.29)	2.03	(0.28)	2.01	(0.28)	1.38	(0.28)	1.32	(0.28)	1.17	(0.27)
7	1.66	(0.31)	1.67	(0.30)	1.90	(0.28)	1.28	(0.29)	1.31	(0.28)	1.46	(0.26)
8	1.85	(0.31)	1.78	(0.30)	1.88	(0.31)	1.10	(0.29)	1.05	(0.29)	1.16	(0.28)
9	1.25	(0.33)	1.33	(0.32)	0.79	(0.32)	0.58	(0.29)	0.67	(0.29)	0.13	(0.29)
10	0.77	(0.35)	0.68	(0.33)	1.25	(0.35)	0.02	(0.31)	−0.02	(0.30)	0.41	(0.31)
Overall	0.94	(0.09)	0.95	(0.09)	0.96	(0.09)	0.61	(0.08)	0.62	(0.08)	0.61	(0.08)
	B: 5th Grade ITBS (1998 vs. 1996), scale score units											
Decile in 1994 3rd-grade achievement distribution												
1	−0.017	(0.023)	−0.015	(0.023)	−0.026	(0.023)	−0.106	(0.032)	−0.105	(0.032)	−0.110	(0.031)
2	0.062	(0.022)	0.063	(0.022)	0.072	(0.022)	0.037	(0.031)	0.037	(0.030)	0.040	(0.030)
3	0.089	(0.021)	0.093	(0.021)	0.081	(0.021)	0.066	(0.030)	0.070	(0.030)	0.053	(0.029)
4	0.114	(0.021)	0.118	(0.020)	0.122	(0.019)	0.095	(0.029)	0.099	(0.028)	0.109	(0.027)
5	0.119	(0.020)	0.119	(0.019)	0.120	(0.019)	0.116	(0.028)	0.117	(0.027)	0.113	(0.027)
6	0.101	(0.020)	0.105	(0.020)	0.094	(0.019)	0.081	(0.027)	0.077	(0.027)	0.084	(0.026)
7	0.053	(0.019)	0.060	(0.019)	0.057	(0.017)	0.115	(0.026)	0.124	(0.026)	0.101	(0.024)
8	0.085	(0.020)	0.095	(0.020)	0.087	(0.019)	0.066	(0.027)	0.075	(0.027)	0.074	(0.026)
9	0.070	(0.019)	0.070	(0.019)	0.038	(0.019)	0.052	(0.029)	0.050	(0.028)	0.000	(0.028)
10	−0.008	(0.019)	−0.006	(0.019)	0.022	(0.019)	−0.081	(0.033)	−0.084	(0.032)	−0.032	(0.033)
Overall	0.066	(0.006)	0.069	(0.006)	0.066	(0.006)	0.043	(0.009)	0.046	(0.009)	0.043	(0.009)
	C: 6th Grade ITBS (1998 vs. 1996), scale score units											
Decile in 1994 4th-grade achievement distribution												
1	0.073	(0.025)	0.064	(0.025)	0.070	(0.024)	0.060	(0.034)	0.054	(0.033)	0.048	(0.032)
2	0.192	(0.025)	0.172	(0.024)	0.196	(0.024)	0.191	(0.034)	0.174	(0.033)	0.201	(0.035)
3	0.235	(0.023)	0.220	(0.023)	0.222	(0.021)	0.233	(0.031)	0.234	(0.032)	0.237	(0.029)
4	0.246	(0.022)	0.237	(0.022)	0.262	(0.020)	0.230	(0.030)	0.226	(0.030)	0.217	(0.028)
5	0.209	(0.021)	0.212	(0.021)	0.209	(0.020)	0.158	(0.029)	0.164	(0.029)	0.172	(0.027)
6	0.259	(0.021)	0.252	(0.021)	0.245	(0.019)	0.186	(0.027)	0.181	(0.027)	0.177	(0.027)
7	0.208	(0.020)	0.205	(0.020)	0.224	(0.018)	0.150	(0.026)	0.149	(0.026)	0.161	(0.023)
8	0.189	(0.019)	0.187	(0.019)	0.187	(0.019)	0.160	(0.026)	0.164	(0.026)	0.158	(0.026)
9	0.152	(0.018)	0.144	(0.018)	0.128	(0.018)	0.049	(0.026)	0.052	(0.026)	−0.002	(0.026)
10	0.086	(0.018)	0.095	(0.017)	0.116	(0.017)	0.037	(0.032)	0.053	(0.031)	0.092	(0.031)
Overall	0.183	(0.007)	0.177	(0.007)	0.184	(0.006)	0.142	(0.009)	0.142	(0.009)	0.143	(0.009)

Note: We describe our estimation procedure in section IV. Table A1 describes the samples by decile. The scale for the ISAT scores ranges from 120 to 200. The ISAT is designed to have a standard deviation of 15 for the population of fifth-grade students in Illinois. The ITBS scores are in grade-equivalent units; for example, 5.1 is the achievement level associated with the end of the first month of fifth grade. Note 17 describes how we calculate the standard errors in our main specification and the specification with additional controls for race, gender, and free-lunch status. We use 1,000 bootstrap replications to compute the standard errors for the models with school fixed effects because we cannot consistently estimate the variance-covariance matrices for regressions that include over 400 school fixed effects.

within broader ability cells while including school fixed effects. Using only within-school variation in student outcomes, we find results that follow the same pattern as those presented in figures 1A and 1B.

For each specification, table 1 also presents the estimated average score gains for the entire sample. We find that NCLB is associated with increases in overall average scores. Thus, our results are consistent with the large body of research that finds positive impacts of accountability systems on average test scores at the state, district, or school level. However, in contrast to this literature, our primary concern is not the extent to which these average gains

generalize to alternative assessments. We emphasize that whatever permanent skill increases are associated with these average gains are not gains enjoyed by the students who are at the bottom of the baseline achievement distribution.[21]

Figures 1A to 1C provide only indirect support for our model because we do not have direct measures of teacher effort, and other mechanisms could generate the patterns we observe in these figures. If schools, in response to NCLB, picked curricula that worked best for students near proficiency and less well for the most and least able students, a similar pattern might emerge. Nonetheless, any alternative explanation for our results must explain how NCLB leads to changes in educational practice that benefit many students but not students with the lowest levels of prior achievement.

Without arbitrary assumptions about the exact shape of the penalty function, our model cannot generate clear predictions concerning exactly how the shape of figures 1A and 1B should change when we restrict the sample to schools that have certain types of baseline students. However, two features of the model are clear. First, students near the proficiency standard ex ante always receive extra attention because this is the most cost-effective strategy for increasing proficiency counts. Second, schools with low ex ante proficiency rates and few students near the proficiency threshold cannot avoid sanctions by simply directing attention to students near the proficiency standard. Thus, if the penalty function is convex enough, these schools will find it optimal to direct some extra effort toward students who are well below the proficiency standard. Although our sample sizes are not large enough to make strong inferences, we find suggestive evidence that students who are far below proficiency do fare better in schools with the lowest expected proficiency rates.

When we repeat our analyses within samples of schools that are comparable in terms of their expected proficiency rates prior to NCLB, we find, as we expect given our model, clear and noteworthy gains among students in the middle deciles of baseline achievement regardless of whether schools are under modest or great pressure from NCLB's AYP rules. Further, there is suggestive but not statistically significant evidence that students at the bottom of the achievement distribution do in fact fare better if they attend a school with expected proficiency rates of less than 25% than if they attend schools with expected proficiency rates between 25% and 40%. Among schools with expected proficiency rates less than 25%, our estimated treatment effects for the bottom two deciles are −0.17 and 0.13 in math and 0.38 and 0.52 in reading, but among schools with expected proficiency rates between 25% and 40%, the corresponding results are −0.77 and −0.48 in math and −0.70 and −0.19 in reading.[22] These differences are noteworthy, but because our sample sizes within school type are so much smaller, we cannot reject the null hypothesis that the effects of NCLB among students in the bottom deciles are the same across these two school types. Nonetheless, the qualitative pattern in these results is consistent with our expectations given our model.[23]

C. *Effects of the 1996 CPS Reforms*

Figures 3A and 3B present estimates of the effects of the 1996 CPS reforms on reading and math scores in fifth grade. Here, we are comparing the performance of students tested in 1998 with the performance that we would have expected from similar students in 1996. The results for fifth-grade reading in figure 3A represent the effects of policy changes that most closely resemble NCLB. Although CPS put reading first in its reform effort and made school-level probation decisions based primarily on proficiency counts in reading, CPS also published school ratings for math and reading in local newspapers based on proficiency counts. Further, fifth graders did not face a threat of summer school if they did poorly on the ITBS, and thus the CPS efforts to end social promotion, which are not part of NCLB, should not have affected results for fifth graders to the same degree that they affected the performance of students in sixth grade. Fifth-grade teachers and parents may well have responded to the promotion hurdles that awaited these students as sixth graders in 1999. However, we do not expect fifth-grade students to make significant changes in their focus and effort based on the consequences attached to sixth-grade exams because children discount the future heavily at this age. This creates an important difference between our fifth- and sixth-grade results.[24] In a standard model of student effort, students will increase their effort in response to an immediate threat of summer school if the cost of such an increase is offset by a significant reduction in the likelihood of attending summer school, and we will see that our results for sixth graders are consistent with this hypothesis.

[21] The results in figures 1A and 1B are also robust to different methods of measuring the heterogeneous effects of NCLB on test scores. We have experimented with finer partitions of the baseline achievement distribution and have used local linear regression methods to estimate the plots in our figures as continuous functions. We also examined numerous mean differences in pre- and postreform test scores for samples of students grouped according to cells defined by both their reading and math scores. Regardless of the methods we have used, we have found no evidence of gains in math or reading scores among students at the bottom of the third-grade achievement distribution, and this is also true regarding our analyses of changes in fifth-grade scores following the 1996 reforms with CPS. Further, we have always found groups of students in the middle of the third-grade achievement distribution who experienced increases relative to prereform baselines in their fifth grade scores.

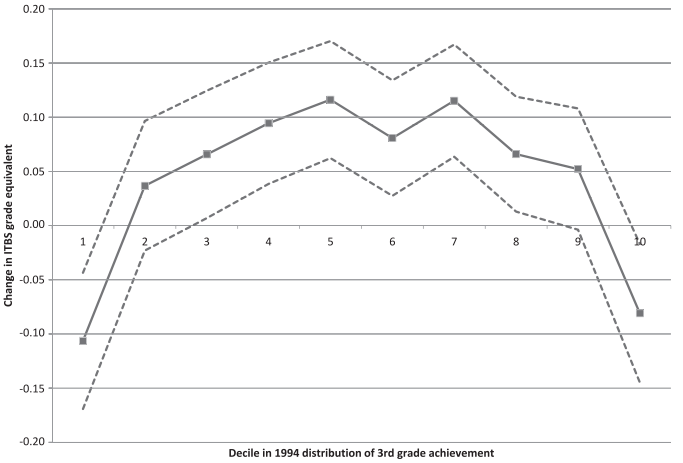
[22] In our model, no student should ever be harmed directly by the introduction of an accountability system because we have made the strong assumption that districts perfectly monitor some baseline level of effort before and after the introduction of accountability, and we do not model group instruction or related choices concerning curricular selection or the pace of instruction. Nonetheless, if schools responded to NCLB by tailoring all group instruction to the needs of students near the proficiency standard, other students could be harmed directly.

[23] In schools with an expected proficiency rate greater than 0.4, our samples of students in the bottom deciles are too small to support meaningful inferences.

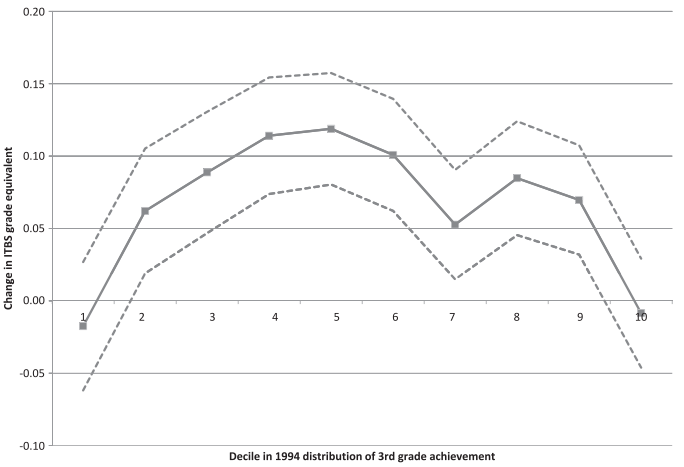
[24] We do not analyze seventh-grade scores in 1998 because the sixth-grade promotion hurdle in 1997 is a source of endogenous composition changes in the 1998 seventh-grade sample.

FIGURE 3.—FIFTH-GRADE RESULTS FOR 1998

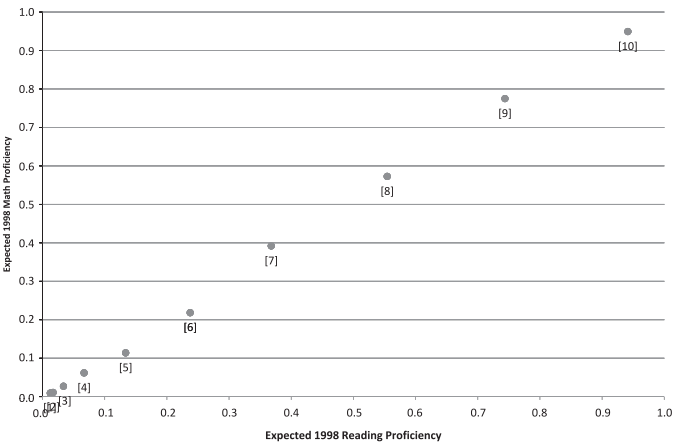
A. CHANGE IN FIFTH-GRADE READING SCORES, 1998 VERSUS 1996



B. CHANGE IN FIFTH-GRADE MATH SCORES, 1998 VERSUS 1996



C. EXPECTED 1998 PROFICIENCY IN FIFTH-GRADE BY DECILE OF THE THIRD-GRADE ACHIEVEMENT DISTRIBUTION FOR 1994



The pattern of results in figure 3A is quite similar to the pattern observed in our analyses of NCLB. Here, the scale is in grade equivalents, and a 0.1 change represents roughly one month of additional achievement. The overall standard deviations of fifth-grade scores in our 2002 samples are roughly 1.2 for math and 1.5 for reading. Thus, estimated achievement gains of 0.1 or slightly more for several cells in the middle of the ability distribution are noteworthy. Still, we find 0 or negative estimated achievement effects among students in either tail. Further, figure 3B shows a similar but slightly less dramatic pattern of changes in fifth-grade math scores. Figure 3C shows that the CPS proficiency standards were slightly more demanding than the ISAT proficiency standards used in 2002, and thus, it is noteworthy that fifth-grade ITBS scores did increase among students in the third decile of the prior achievement distribution, even though one would have expected less than 5% of these students to pass either the math or reading thresholds in the prereform period. Nonetheless, the teachers and parents of these students knew that they would face a promotion hurdle as sixth graders in 1999, and as we demonstrate below, the standards for promotion were within the reach of these students.

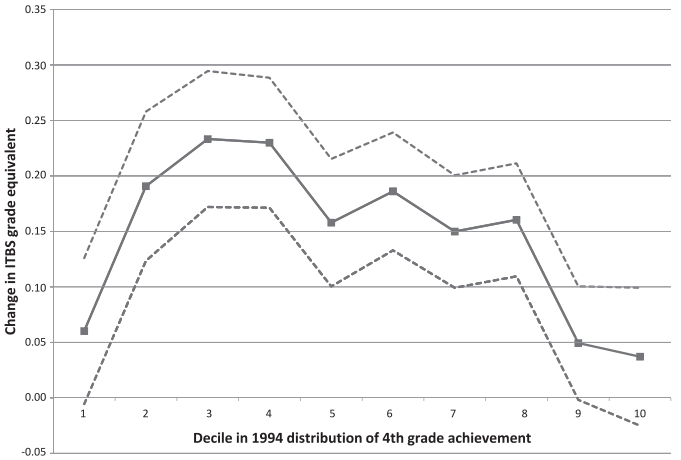
Figures 4A and 4B present results for sixth graders. Here, we are clearly not measuring just the effects of the school probation rules and the public reporting of proficiency counts in local newspapers. We anticipate that the rules governing summer school attendance and retention decisions shaped not only the actions of teachers and parents but also the effort of students during the school year. Students in sixth grade faced summer school if they performed below certain targets in reading or math, and these targets were much lower than the proficiency standards used to measure school performance. Taking all of these factors into account, we are not surprised that while our results for sixth graders follow the same overall pattern observed among fifth graders, the estimated sixth-grade gains associated with the CPS reforms are larger at every decile in both math and reading, and there is some evidence of gains even in the lowest decile.

Figure 4C is similar to figure 3C except that it plots, for each decile, the probabilities of exceeding the summer school cutoffs for sixth graders.[25] The striking difference between figures 3C and 4C may offer some insight concerning why estimated gains from accountability in figures 4A and 4B are more apparent in the lower deciles of the achievement distribution. Even students in the lowest decile of fourth-grade achievement had almost a 20% chance of reaching the individual math or reading cutoffs that determined summer school attendance after sixth grade, and White and Rosenbaum (2007) suggest that among sixth

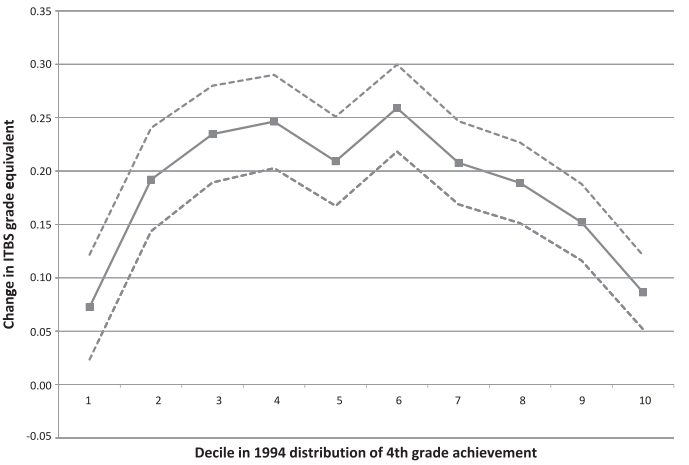
[25] The primary focus of Jacob (2005) is the average change in scores in response to the introduction of the CPS accountability system. However, Jacob does examine an interaction between student ability and the impact of high-stakes testing. Although Jacob's method involves using scores from many years and thus many different forms of the ITBS exam as well as a much more restrictive specification of how heterogeneity influences the impacts of high stakes, he comes to a conclusion that squares broadly with our results: "Students who had been scoring at the 10th–50th percentile (in the national distribution) in the past fared better than their classmates who had either scored below the 10th percentile, or above the 50th percentile" (pp. 776–777).

FIGURE 4.—SIXTH-GRADE RESULTS FOR 1998

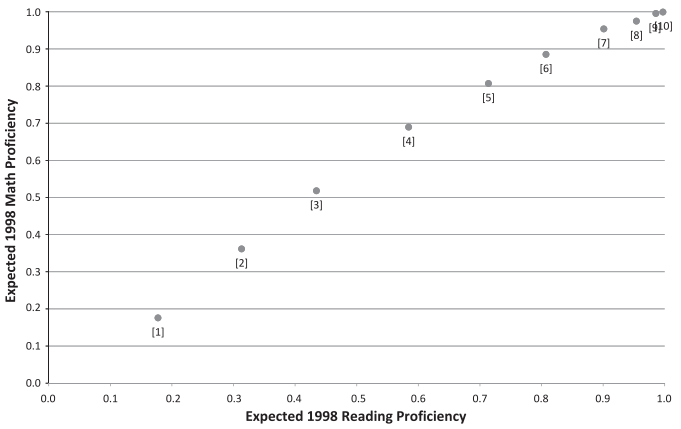
A. CHANGE IN SIXTH-GRADE READING SCORES, 1998 VERSUS 1996



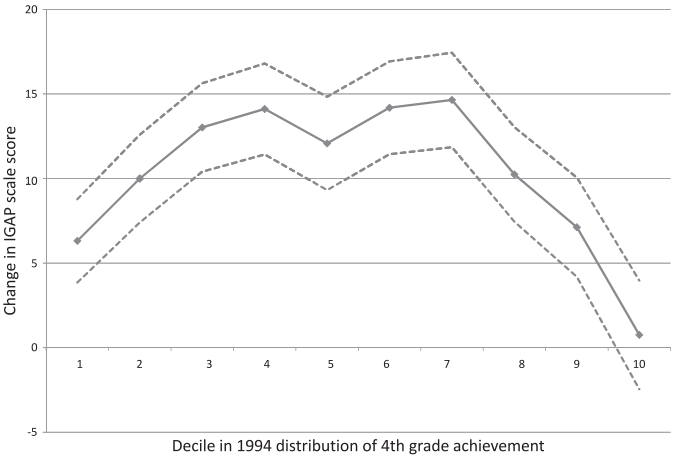
B. CHANGE IN SIXTH-GRADE MATH SCORES, 1998 VERSUS 1996



C. EXPECTED 1998 PASS RATES IN SIXTH GRADE: SUMMER SCHOOL CUTOFFS BY DECILE OF THE FOURTH-GRADE ACHIEVEMENT DISTRIBUTION FOR 1994



D. CHANGE IN SIXTH-GRADE MATH SCORES ON LOW-STAKES IGAP TEST, 1998 VERSUS 1996



graders, CPS schools targeted their instructional efforts toward students who could avoid summer school only if they made progress during the school year.[26] We argue in section II that less demanding proficiency targets can increase the amount of attention that teachers devote to less able students, and the contrast between our results for fifth and sixth grades is consistent with our conjecture. However, we cannot rule out the possibility that even students at the lowest levels of prior achievement simply worked harder than similar students in previous cohorts because they wanted to avoid summer school.

Figures 4A and 4B indicate that students in the third and fourth deciles of prior achievement scores scored over 0.2 higher in math and reading than one would have expected prior to the 1996 reforms. These are large effects since 0.2 represents two full months of achievement on the ITBS grade-equivalent scale, and it is worth noting that figure 4C implies that CPS set the summer school cutoff scores such that students in these deciles faced both a significant chance of avoiding summer school as well as a significant chance of attending summer school depending on how they progressed during the year.[27]

As we note above, table 1 contains results from two robustness checks that we have conducted for each of our analyses. These results come from models of ITBS achievement that included school fixed effects and models that

[26] However, it is not completely clear that the least able CPS students benefited from this program. In a previous version of the paper, we presented results for twenty prior achievement cells. The estimated sixth-grade effects for those in the bottom 5% of the ability distribution were quite close to zero and not statistically significant. See Roderick and Engel (2001) for more work on the motivational responses of low-achieving children to the retention policy in CPS.

[27] Another literature explores how students respond when they face different stakes and performance standards on tests. See Betts and Grogger (2003), as well as Becker and Rosen (1992), who apply insights from Lazear and Rosen's (1981) tournament model to the design of academic testing systems that determine rewards and punishments for students. This literature suggests that less able students will not be affected by these systems if they have no realistic chance of ever reaching these key cutoff scores. Thus, the decision by CPS to set modest standards for grade promotion may have been advantageous for generating more effort among low-achieving students.

TABLE 2.—TREATMENT EFFECT ON PROFICIENCY RATES

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Math				Reading			
	Main Effect	SE	With Demographic Controls	SE	Main Effect	SE	With Demographic Controls	SE
	A: Proficiency on 5th-Grade ISAT (2002 vs. 2001)							
Decile in 1999 3rd-grade achievement distribution								
1	0.006	(0.004)	0.007	(0.004)	−0.002	(0.005)	−0.001	(0.005)
2	0.014	(0.006)	0.016	(0.006)	−0.006	(0.008)	−0.006	(0.008)
3	0.057	(0.009)	0.056	(0.009)	0.019	(0.011)	0.018	(0.011)
4	0.077	(0.013)	0.077	(0.012)	0.041	(0.014)	0.038	(0.014)
5	0.082	(0.015)	0.077	(0.014)	0.035	(0.015)	0.032	(0.015)
6	0.114	(0.017)	0.111	(0.016)	0.076	(0.017)	0.073	(0.017)
7	0.080	(0.016)	0.080	(0.016)	0.078	(0.016)	0.078	(0.016)
8	0.054	(0.015)	0.053	(0.015)	0.077	(0.014)	0.075	(0.014)
9	0.008	(0.012)	0.010	(0.012)	0.031	(0.010)	0.033	(0.010)
10	0.011	(0.006)	0.010	(0.006)	0.016	(0.006)	0.015	(0.006)
Overall	0.047	(0.004)	0.047	(0.003)	0.033	(0.004)	0.032	(0.004)
	B: Proficiency on 5th-Grade ITBS (1998 vs. 1996)							
Decile in 1994 3rd-grade achievement distribution								
1	0.002	(0.003)	0.002	(0.003)	0.001	(0.004)	0.002	(0.004)
2	0.008	(0.004)	0.008	(0.004)	0.004	(0.004)	0.004	(0.004)
3	0.013	(0.006)	0.014	(0.006)	0.012	(0.006)	0.012	(0.006)
4	0.026	(0.008)	0.027	(0.008)	0.030	(0.009)	0.030	(0.009)
5	0.032	(0.011)	0.032	(0.010)	0.033	(0.011)	0.038	(0.011)
6	0.042	(0.013)	0.045	(0.013)	0.034	(0.013)	0.031	(0.013)
7	0.031	(0.014)	0.034	(0.014)	0.052	(0.014)	0.056	(0.014)
8	0.045	(0.014)	0.049	(0.014)	0.013	(0.015)	0.016	(0.015)
9	0.037	(0.011)	0.037	(0.011)	0.027	(0.013)	0.027	(0.013)
10	−0.006	(0.006)	−0.005	(0.006)	−0.011	(0.007)	−0.010	(0.007)
Overall	0.023	(0.003)	0.025	(0.003)	0.020	(0.003)	0.021	(0.003)
	C: Proficiency on 6th-Grade ITBS (1998 vs. 1996) Relative to Summer School Cutoff							
Decile in 1994 4th-grade achievement distribution								
1	0.054	(0.011)	0.051	(0.011)	0.037	(0.011)	0.037	(0.012)
2	0.101	(0.014)	0.093	(0.014)	0.072	(0.015)	0.064	(0.015)
3	0.119	(0.014)	0.113	(0.014)	0.111	(0.015)	0.113	(0.015)
4	0.084	(0.012)	0.081	(0.012)	0.081	(0.014)	0.080	(0.014)
5	0.054	(0.010)	0.057	(0.010)	0.047	(0.013)	0.050	(0.013)
6	0.047	(0.008)	0.044	(0.008)	0.059	(0.011)	0.057	(0.010)
7	0.015	(0.006)	0.014	(0.006)	0.030	(0.008)	0.031	(0.008)
8	0.012	(0.004)	0.012	(0.004)	0.014	(0.006)	0.014	(0.006)
9	0.002	(0.002)	0.002	(0.003)	0.008	(0.003)	0.008	(0.003)
10	−0.001	(0.001)	−0.001	(0.004)	0.001	(0.002)	0.001	(0.004)
Overall	0.047	(0.003)	0.049	(0.003)	0.044	(0.003)	0.044	(0.003)

Note: See notes to table 1. This table presents results that parallel those in table 1, but the dependent variable is an indicator for being proficient in fifth grade or exceeding the summer school cutoff in sixth grade. Further, we use a logit model rather than linear regression to create predicted proficiency or pass rates given student characteristics. We calculate the standard errors using 1,000 bootstrap replications.

included additional controls for race, gender, and free-lunch eligibility. Results from these alternative specifications are quite similar to those in figures 3A, 3B, 4A, and 4B.

Table 2 presents results that parallel those in table 1, but here the dependent variables are indicators for scoring above either proficiency standards or the cutoff scores associated with the summer school policy. The results follow the patterns that we expect given our figures. However, some readers may wonder why the changes in summer school pass rates are so modest among sixth graders in deciles 6 through 8 given that figures 4A and 4B present noteworthy gains in achievement for these deciles. The answer lies in figure 4C. Because the summer school cutoff scores are quite low relative to the NCLB or CPS proficiency standards, the vast majority of students in these deciles should have been able to avoid summer school

without receiving extra help from their teachers or working harder on their own. For these students, extra help from teachers or extra effort in class provided insurance against a small but not negligible risk of summer school. Such actions could easily produce noteworthy achievement gains while having small impacts on average pass rates.

D. Changes in a Low-Stakes Outcome

We note above that much of the literature on responses to high-stakes testing programs addresses the possibility that teachers take actions that inflate students' scores on high-stakes exams relative to their actual skill levels. Thus, some may wonder how the reforms that we analyze affect distributions of outcomes on low-stakes assessments. We are not able to address this issue in detail because almost all of the assessments that students took outside these two accountability systems are either not comparable for the pre- and postreform cohorts or not given in the grades we consider.[28]

However, we are able to construct figure 4D. This figure parallels figure 4B, but the outcome variable is the sixth-grade math IGAP (Illinois Goals Assessment Program) test and not the sixth-grade math ITBS test used in the CPS accountability plan. The cells are defined exactly as in figure 4B. As before, only students who took the ITBS in both fourth and sixth grades are included in the analyses, and the conditioning variables are the fourth-grade ITBS scores.

Three features of the IGAP results are noteworthy. First, because the standard deviation of sixth-grade math IGAP scores is 86, the gains reported in deciles 3 through 7 of figure 4D are noteworthy and all more than .1 standard deviations. Second, the gains in figure 4D are typically a bit smaller than those in figure 4B if both sets of results are expressed in standard deviation units.[29] Third, the same hump-shape pattern that appears in figure 4B also appears in 4D. These results provide additional evidence that the CPS system benefited students in the middle of the achievement distribution more than those at the bottom of the distribution.

We also note that several of the estimated gains in IGAP scores presented in figure 4D are almost certainly inflated by the process of selection into IGAP test taking. The samples in figures 4A and 4B include sixth graders in the 1998 and 1996 cohorts who took the ITBS in fourth and sixth grades, but students included in the analyses for figure 4D must also take the sixth-grade IGAP math test in 1996 or 1998. In both 1996 and 1998, those who take the IGAP have higher average baseline achievement. Further, if we restrict our attention to students who take both the IGAP and ITBS, we find larger implied gains in ITBS math scores between 1998 and 1996 than those presented for the full sample in figure 4B. For most deciles, the differences between our results in figure 4B and those from parallel analyses restricted to students who take both the ITBS and IGAP are quite small. However, this is not the case in the first and second deciles, where, respectively, over one-fourth and one-tenth of the students represented in figure 4B did not take the IGAP.[30]

To gauge how much selection into IGAP testing affects the results in figure 4D, we calculate a simple selection correction factor based on the results in figure 4B and our parallel set of ITBS math results for the sample of students who took both the IGAP and ITBS. We calculate ratios of the estimated changes in ITBS math scores presented in figure 4B to the corresponding estimated changes among the select sample of students who took both IGAP and ITBS. We then form selection-corrected IGAP results by taking the product of these ratios and our estimated IGAP effects in figure 4D.[31] The implied corrections are small for most deciles. However, the estimated IGAP gain for decile 1 falls from 6.3 to 2.9, and the estimated gain for decile 2 falls from 10 to 8.7. These selection corrections do not change the pattern of results presented in figure 4D but simply accentuate the overall hump shape that is already present.

E. Summary

Our results show a consistent pattern across assessments, subjects, and years. Accountability systems built around proficiency counts do not generate a uniform distribution of changes in measured achievement. Students who have no realistic chance of becoming proficient in the near term appear to gain little from the introduction of these systems. However, our results from schools with ex ante proficiency rates of less than 25% provide suggestive evidence that low-achieving students may fare slightly better if they attend schools that cannot meet target proficiency levels by concentrating only on students who are already near proficiency. Our results for sixth graders in 1998, who were able to avoid summer school if they scored above relatively modest thresholds, are consistent with the proposition that

[28] The IGAP reading is not comparably scored over time. See Jacob (2005) for details. Further, the IGAP was not given in fifth grade. By the time of NCLB in 2002, the ITBS had actually become a relatively low-stakes exam in Chicago. However, between 2001 and 2002, CPS changed to a completely different form of the exam.

[29] The standard deviation of sixth-grade ITBS scores is 1.38. The average ITBS gain is .132 standard deviations, while the average IGAP gain is 0.118 standard deviations. Given the correction for selection into IGAP testing that we describe in the following paragraph, the average IGAP gain falls to 0.11 standard deviations. Using regression models with sixth-grade ITBS and IGAP scores from 1994 through 1998 as dependent variables and student demographic characteristics, policy indicators, and time trends as control variables, Jacob (2005, pp. 781–782) reaches a similar conclusion concerning the impact of the CPS reform on sixth-grade ITBS math gains. However, he cautiously attributes the growth in sixth-grade IGAP math scores to a preexisting trend in IGAP achievement.

[30] We have constructed figure 4B while restricting our outcome samples to students with valid IGAP math scores, and we find that the estimated gains in ITBS math scores more than double for the first decile and increase by roughly 13% for the second.

[31] This method is valid if the ratios of treatment effects for the select versus full samples are the same for IGAP and ITBS effects. It makes sense that this correction should matter most in the first decile. Rates of IGAP testing are lowest in this decile. Further, those who did not take the IGAP enjoy lower baseline achievement and are less likely to view the summer school cutoffs as realistic goals.

lower proficiency levels shift the benefits of these systems toward students with lower baseline achievement levels.[32]

V. Potential Reforms to Accountability Systems

The central lesson of the model and empirical work presented here is that an accountability system built around proficiency counts may not help students who are currently far above or far below these thresholds. In this section, we ask whether recent proposals for changing the AYP system can help make NCLB a policy that generates improved instruction for all students. We assume that the goal of NCLB or related accountability programs is to induce a uniform increase in the amount of extra instruction that teachers give to students of all abilities. We acknowledge that there is no reason to believe that increasing the effort allocated to each student by the same amount is socially optimal. However, by analyzing how different AYP scoring systems influence the allocation of teacher effort relative to this standard, we illustrate the key issues that designers of accountability systems must face when trying to elicit any particular distribution of effort that may be deemed desirable.

Education policymakers are currently devoting significant attention to two alternative schemes for measuring AYP at the school level. First, several states have adopted indexing systems based on multiple thresholds.[33] In such a system, students who score above the highest threshold contribute, as in other states, one passing score toward the school's proficiency count. However, students who fall short of this highest threshold but do manage to exceed lower thresholds count as varying fractions of a passing student depending on how many thresholds they meet. Other states have adopted value-added systems that measure how much scores have improved, on average, between two test dates.[34]

These approaches do not build in strong incentives to focus attention only on students near a single proficiency standard, and one can easily construct examples in which these systems will mitigate the number of students who receive no extra attention under NCLB. However, there are important trade-offs between the two approaches.

Using the notation from section III, consider the following index system:

$$\min_{e_i} \mathrm{I}\left[\frac{1}{T^u}\sum_{i=1}^{N} E(t_i)\right] + \sum_{i=1}^{N} c(e_i). \tag{2}$$

T^u is the maximum possible score on the high-stakes assessment, and we normalize the floor of the scale to 0. Here,

$$t_i = \min[(T^u, \max(0, t_i = e_i + \alpha_i + \varepsilon_i)].$$

Thus, all scores are constrained to be between the floor and ceiling of the scale used for assessment: $t_i \in [0, T^u]\ \forall\ i = 1, 2, \ldots, N$. In a complete analysis of equation (2), we would need to address the fact that the relationship between e_i and the expected value of student i's test score may be a function of both the floor and ceiling on the test scale. However, we assume that the tests in question are designed to ensure that neither the floor nor ceiling on the scale is relevant for investment decisions regarding students, regardless of student ability.[35]

In the index system described by equation (2), proficiency for a given student is no longer 0 or 1 but rather the student's score expressed as a fraction of the maximum score, and the penalty function $\mathrm{I}[\cdot]$ describes how sanctions decline, for a school of size N, as the total proficiency count of the school increases. This indexing system resembles those used in some states, but it differs in two ways. First, the indexing is continuous, so that all score increases count the same toward the school's proficiency score regardless of a student's initial ability, α_i.[36] Second, because we have set the standard for proficiency at the maximum possible score and assumed this score is never reached, we have eliminated the existence of students for whom $e_i = 0$ simply because they are already too accomplished relative to the proficiency standard.

We can easily compare this characterization of indexing in equation (2) to the following value-added system:

$$\min_{e_i} \Gamma\left[\sum_{i=1}^{N} (E(t_i) - \bar{\alpha})\right] + \sum_{i=1}^{N} c(e_i). \tag{3}$$

Here, $\bar{\alpha}$ is the average performance in the school on a previous assessment. Because of the linearity in our model, equation (3) describes a value-added system in which schools of a given size are rewarded or sanctioned according to $\Gamma(\cdot)$ based on their total net improvement in student achievement. With regard to the goal of leaving no child behind, both of the systems in equations (2) and (3) repre-

[32] We have conducted our analyses of fifth graders using proficiency as the outcome variable, and as one would expect based the results we report in our figures, proficiency rates in the bottom two deciles of achievement remain almost constant following the two reforms. We do see small increases in the number of 1998 sixth graders in the first decile who meet the summer school cutoffs for reading and math achievement. This is expected because the cutoffs are set at such low levels.

[33] As of spring 2007, these are Alabama, Florida, Iowa, Louisiana, Massachusetts, Michigan, Minnesota, Mississippi, New Hampshire, New Mexico, Oklahoma, Pennsylvania, Rhode Island, South Carolina, Vermont, Wisconsin, and Wyoming. New York also has a small indexing component in their system.

[34] As of January 2009, fifteen states had received federal approval of growth model plans: North Carolina, Tennessee, Delaware, Arkansas, Florida, Iowa, Ohio, Alaska, Arizona, Minnesota, Missouri, Pennsylvania, and Texas.

[35] The existence of floors or ceilings implies that there may exist regions of ability types at the top and bottom of the ability distribution such that the return to investment in students is diminished because their observed scores are likely to remain at the ceiling or floor even if their latent scores improve. By ignoring these possibilities, we are implicitly assuming that the distribution of ability types and the distribution of measurement errors are bounded in a manner that makes the floor and ceiling scores unattainable given the optimal vector of effort choices.

[36] We do not think of ability as a fixed endowment but rather as the level of competency at the beginning of a given school year, which should reflect investments made by both schools and parents in previous years.

sent the best of all possible worlds in many respects. Any given increase in the expected score of any given student makes the same contribution to the school's standing under the accountability system regardless of the student's initial achievement level. Therefore, in this setting, the optimal vector of effort allocations will dictate an identical increase in attention for all students as long as we maintain our assumption that $c(e)$ is strictly convex.

The systems described here are useful benchmarks because they demonstrate what is required to design an accountability system that does not direct effort toward a particular group of students. Note that the systems described here require a team of incredibly skilled test developers. The test scales in equations (2) and (3) are such that the effort cost of increasing an individual's expected score by any fixed amount is the same for all students. In practice, differences in the costs of improving student scores by particular increments at different points on a given scale will influence the allocation of effort among students.[37]

While both indexing and value-added systems offer means for eliciting improved allocations of teacher effort to all students and not just those near proficiency standards, indexing and value-added are not equally desirable on all dimensions. Under any system that ties rewards and sanctions to levels of achievement, including a continuous indexing system, the minimized sum of effort costs and sanctions borne by the staff is a function of the distribution of prior student achievement in the school. Thus, it may be difficult to design an index system that challenges the best schools without setting goals for disadvantaged schools that are not attainable given their resources. The Clotfelter et al. (2004) results suggest that when indexing systems set unattainable goals for schools in disadvantaged communities, these systems may actually do harm by causing these schools to lose the teachers they need most. Under the value-added system described in equation (3), the total cost of achieving the optimal proficiency score is not affected by the distribution of initial ability.[38] Thus, it might be possible to use such a value-added system to increase the quality of instruction for all students without distorting the supply of teachers among schools.[39]

Still, value-added methods are not a panacea. To begin, value-added measures are often much noisier than measures of the current level of student performance. In principle, one could address this concern by developing more reliable assessments, but it is still important to note that teachers may well demand increases in other aspects of their compensation if their standing under an accountability system is greatly influenced by the measurement error in performance measures.[40]

In addition, the absence of a natural scale for knowledge raises important questions about any method that seeks to determine which groups of students made the most academic progress. Reardon (2007) uses data from the Early Childhood Longitudinal Study–Kindergarten cohort (ECLS–K) to show that measured differences between the magnitude of the black-white test score gap in first grade and fifth grade among a single cohort of students can be quite sensitive to the specific scale used to report the scores, even if all scores from all candidate scales are standardized to have a mean of 0 and a variance of 1. The ECLS–K data do not permit researchers to make definitive statements about how much bigger the black-white test score gap is among fifth graders than among first graders because black and white students begin first grade with different achievement levels and there is no natural metric for knowledge that tells us how to compare the size of this achievement gap with the corresponding gap observed in fifth grade. Since value-added measures are measures of achievement growth for a population of students, the claim that value-added is greater in school A than school B is a claim that, on average, achievement growth was greater in school A than school B during the past year. But if it is difficult to make robust judgments concerning whether achievement growth was greater among white students than black students in a nationally representative panel, it will also be difficult to make robust judgments concerning the relative magnitudes of average achievement growth in different schools.[41]

In the end, designers of accountability systems face an important trade-off. Any index system built around cutoff scores will make it more costly to attract teachers to teach in disadvantaged schools as long as all schools are held to the same proficiency standards. On the other hand, systems built around measures of achievement growth will provide incentives and hand out sanctions based on performance measures that may be noisy and not robust to seemingly arbitrary choices concerning scales. Nonetheless, both methods may reduce the incentives some schools currently face to leave the least advantaged behind.

[37] Further, if $c(e)$ is linear instead of strictly convex, it is easy to construct examples such that the optimal vector of effort allocations for both of the problems above includes increased attention for only some students. Given a penalty function, there will be a specific total sum of test scores or a total proficiency score such that the constant marginal cost of raising the total beyond this point is greater than the reduction in sanctions associated with such an increase, and there is nothing in the structure of this problem that guarantees $e_i > 0$ for all i at this point. Thus, even with these ideal scales, we need to assume strictly decreasing returns to teacher effort at the student level to rule out the possibility that schools will target only a subset of students in their efforts to avoid sanctions.

[38] This can be shown easily by substituting the formula for t_i into equation (3), if one assumes that the floor and ceiling on the test scale are never binding at the optimal effort vector.

[39] Here, we are implicitly assuming that the increase in the effort cost of teaching will not generate a decline in teacher quality that offsets the increased effort given by remaining teachers.

[40] See Kane and Staiger (2002) for more on problems caused by measurement error in value-added systems.

[41] Reardon's (2007) results are driven in part, by the fact that the typical black student is making gains over a different region of the scale than the typical white student. Because students sort among schools on ability, a similar problem arises when measuring relative achievement growth among schools.

VI. Conclusion

A significant ethnographic literature documents instances in specific schools where schools responded to accountability systems by targeting so-called bubble kids for extra help while simultaneously providing no special attention to students who were already proficient or unlikely to become proficient given feasible interventions. Here, we use unique data from Chicago that permit us to cleanly measure how the entire distribution of student achievement changes following the introduction of accountability systems built around proficiency counts. Our findings are quite consistent with the conclusions in the ethnographic literature, and we are the first to document educational triage on a large scale by comparing cohorts of students who took the same exams under different accountability regimes.

Our results do not suggest that NCLB has failed to improve performance among all academically disadvantaged students in Chicago. Figures 1A and 1B show that 2002 ISAT test scores among fifth graders were higher than one would have expected prior to NCLB over most of the prior achievement distribution, and it is important to note that even CPS students in the fourth decile of the third-grade achievement distribution faced just over 20% and just under 15% chances of being proficient in reading and math respectively prior to NCLB. Thus, many low-achieving students in Chicago appear to have done better on ISAT under NCLB than they would have otherwise. However, for at least the bottom 20% of students, there is little evidence of significant gains and a possibility of lower-than-expected scores in math. If we assume that similar results hold for all elementary grades now tested under NCLB, we have reason to believe that at a given time, there are more than 25,000 CPS students being left behind by NCLB.[42]

This large number is the result of several factors interacting together. First, as a state, Illinois has set standards that are challenging for disadvantaged students. According to a report by the Chicago Consortium on School Research, Easton et al. (2003), just over half of the nation's fifth graders would be expected to achieve the ISAT proficiency standard in reading, and just under half would be expected to achieve the ISAT standard in math. Second, students in Chicago are quite disadvantaged. More than 80% of CPS students receive free or reduced-price lunch benefits. Third, CPS is one of the largest districts in the country.

We do not have data on individual test scores from other states, and we cannot assess the extent to which our results from Chicago reflect a pattern that is common among other school districts in other states. However, we have reason to believe that while the pattern of NCLB effects we have identified may not be ubiquitous, it is also not unique to Chicago. New York City, Cincinnati, Cleveland, and many other cities educate large populations of disadvantaged students in states with accountability systems that are roughly comparable to the 2002 system implemented in Illinois.[43] Based on our results, it is reasonable to conjecture that hundreds of thousands of academically disadvantaged students in large cities are currently being left behind because the use of proficiency counts in NCLB does not provide strong incentives for schools to direct more attention toward them. Further, NCLB may be generating this type of educational triage in nonurban districts as well.[44] Any school that views AYP as a binding constraint and also educates a significant number of students who have little hope of reaching proficiency faces a strong incentive to shift attention away from their lowest-achieving students and toward students near proficiency.

Because our results show significant increases in achievement for students near the proficiency standard, our results are consistent with the proposition that accountability systems can generate increases in achievement. However, our results also indicate that rules used to transform the test score outcomes for all students into a single set of accountability ratings for schools play an important role in determining which students experience these achievement gains. More work is required to design systems that truly leave no child behind.

[42] This is a conservative estimate. There are more than 30,000 students in the bottom 20% of the current third- to eighth-grade CPS student population. Although we cannot rule out the possibility that some students in the bottom two deciles of the achievement distribution receive extra help because their teachers see potential that is not reflected in their prior test scores, our results suggest that few students fall into this category. Further, we conjecture that there are as many or more students in the third and fourth deciles who receive little or no extra help because their teachers realize that they are less likely to improve than other students with similar prior achievement levels.

[43] See National Center for Education Statistics (2007). On the other hand, Boston, Detroit, and Philadelphia are in states that use index systems to calculate AYP. Further, Houston, Dallas, and other cities in Texas face a state accountability system built around proficiency standards that are not as demanding as the 2002 standards in Illinois and possibly more in reach for disadvantaged students.

[44] Commercial software now exists that makes it easier for schools to monitor and improve their AYP status. See http://www.schoolnet.com for an example. Schools that wish to create lists of students who are most likely to become proficient given extra instruction can easily do so.

REFERENCES

Becker, William E., and Sherwin Rosen, "The Learning Effect of Assessment Evaluation in High School," *Economics of Education Review* 11:2 (1992), 107–118.

Betts, Julian, and Jeffrey Grogger, "The Impact of Grading Standards on Student Achievement, Educational Attainment, and Entry-Level Earnings," *Economics of Education Review* 22:4 (2003), 343–352.

Booher-Jennings, Jennifer, "Below the Bubble: 'Educational Triage' and the Texas Accountability System," *American Educational Research Journal* 42:2 (2005), 231–268.

Bryk, Anthony S., "No Child Left Behind, Chicago Style," in Paul E. Peterson and Martin R. West (Eds.), *No Child Left Behind? The Politics and Practice of School Accountability* (Washington, DC: Brookings Institution Press, 2003).

Burgess, Simon, Carol Propper, Helen Slater, and Deborah Wilson, "Who Wins and Who Loses from School Accountability? The Distribution of Educational Gain in English Secondary Schools," CMPO working paper (July 2005).

Carnoy, Martin, and Susanna Loeb, "Does External Accountability Affect Student Outcomes? A Cross-State Analysis," *Educational Evaluation and Policy Analysis* 24:4 (2002), 305–331.
Clotfelter, Charles, Helen Ladd, Jacob Vigdor, and Aliaga Diaz, "Do School Accountability Systems Make It More Difficult for Low-Performing Schools to Attract and Retain High-Quality Teachers?" *Journal of Policy Analysis and Management* 23:3 (2004), 251–271.
Cullen, Julie B., and Randall Reback, "Tinkering toward Accolades: School Gaming under a Performance Accountability System," in Timothy J. Gronberg and Dennis W. Jansen (Eds.), *Advances in Applied Microeconomics* 14 (2006), 1–34.
de Vise, Daniel, "Rockville School's Efforts Raise Questions of Test-Prep Ethics," *Washington Post,* March 4, 2007.
Easton, John Q., Macarena Correa, Stuart Luppescu, Hye-Sook Park, Steve Ponisciak, Todd Rosenkranz, and Sue Sporte, "How Do They Compare? ITBS and ISAT Reading and Mathematics in Chicago Public Schools, 1999 to 2002," Consortium for Chicago School Research research data brief (February 2003).
Gillborn, David, and Deborah Youdell, *Rationing Education: Policy, Practice, Reform and Equity* (Philadelphia: Open University Press, 2000).
Grissmer, David, and Ann Flanagan, "Exploring Rapid Achievement Gains in North Carolina and Texas," National Education Goals Panel (November 1998).
Hanushek, Eric A., and Margaret E. Raymond, "Does School Accountability Lead to Improved Student Performance?" NBER working paper no. 10591 (2004).
Holmstrom, Bengt, and Paul Milgrom, "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership and Job Design," *Journal of Law, Economics and Organization* 7 (1991), 24–52.
Jacob, Brian A., "A Closer Look at Achievement Gains under High-Stakes Testing in Chicago," in Paul E. Peterson and Martin R. West (Eds.), *No Child Left Behind? The Politics and Practice of School Accountability* (Washington, DC: Brookings Institution Press, 2003).
——— "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools," *Journal of Public Economics* 89:5–6 (2005), 761–796.
Jacob, Brian A., and Steven Levitt, "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics* 118:3 (2003), 843–877.
Kane, Thomas J., and Douglas O. Staiger, "Volatility in School Test Scores: Implications for Test-Based Accountability Systems," *Brookings Papers on Education Policy* 5 (2002), 235–283.
Koretz, Daniel M., "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity," *Journal of Human Resources* 37:4 (2002), 752–777.
Lazear, Edward P., "Speeding, Terrorism, and Teaching to the Test," *Quarterly Journal of Economics* 121:3 (2006), 1029–1061.
Lazear, Edward P., and Sherwin Rosen, "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy* 89:5 (1981), 841–864.
National Center for Education Statistics, "Mapping 2005 State Proficiency Standards onto the NAEP Scales," NCES report 2007-482 (2007).
Reardon, Sean, "Thirteen Ways of Looking at the Black-White Test Score Gap," Stanford University mimeograph (March 2007).
Reback, Randall, "Teaching to the Rating: School Accountability and the Distribution of Student Achievement," *Journal of Public Economics* 92:5–6 (2008), 1394–1415.
Roderick, Melissa, and Mimi Engel, "The Grasshopper and the Ant: Motivational Responses of Low-Achieving Students to High-Stakes Testing," *Educational Evaluation and Policy Analysis* 23:3 (2001), 197–227.
Springer, Matthew G., "The Influence of an NCLB Accountability Plan on the Distribution of Student Test Score Gains," *Economics of Education Review* 27:5 (2008), 556–563.
Wick, John W., "Independent Assessment of the Technical Characteristics of the Illinois Standards Achievement Test (ISAT)" (Chicago: Illinois State Board of Education, 2003).
White, Katie Weits, and James E. Rosenbaum, "Inside the Black Box of Accountability: How High-Stakes Accountability Alters School Culture and the Classification and Treatment of Students and Teachers," in Alan R. Sadovnik, Jennifer A. O'Day, George W. Bohrnstedt, and Kathryn M. Borman (Eds.), *No Child Left Behind and the Reduction of the Achievement Gap: Sociological Perspectives on Federal Education Policy* (Florence, KY: Routledge, 2007).

APPENDIX A

Data Construction

In our analyses of the effects of NCLB, we restrict our samples to students who were tested in fifth grade in 2002, the first year of NCLB, or 2001. We further restrict the sample to students who were last tested in third grade exactly two years prior. Here, we discuss two alternative procedures.

First, we could have simply selected the first or last third-grade test available for each fifth-grade student in our 2001 and 2002 samples without restricting the sample interval between scores. We chose not to pursue this strategy because the ISAT test was not given until 1999. Students tested in fifth grade in 2001 who entered third grade in 1998 and then either repeated part or all of third grade or fourth grade do not have an ISAT score for their initial third-grade year, and depending on the details of their grade progression, they may not have a third-grade ISAT score at all. This is not true among similar students who entered third grade in 1999 and were tested as fifth graders in 2002. Thus, the sample of fifth graders tested in 2001 with valid third-grade scores contains fewer students who experienced retention problems in third or fourth grade than the comparable sample of fifth graders tested in 2002. By restricting the samples to students who last tested in third grade exactly two years prior, we are holding the progression patterns in the treatment and control samples constant.

A second alternative procedure involves conditioning on a different progression pattern by restricting the samples to students tested two years prior during their first year in third grade. These samples would include only students with "normal" grade progression. We conducted analyses on these samples and found results that are quite similar to those in figures 1A and 1B.

For the final time in 1999, 27,205 students with valid third-grade scores took the ISAT in third grade. The comparable sample for 2000 contains 27,851 students; 20,060 of these 1999 third graders and 21,199 of these 2000 third graders appear in the ISAT fifth-grade test files for 2001 and 2002, respectively. Thus, the sample retention rate is slightly higher in the 2000–2002 sample (73.7% versus 76.1%). One source of this difference in retention rates is that there are fewer student ID number matches looking forward from the 1999 sample. This primarily reflects fewer exits from CPS for the 2000 sample as well as fewer student ID numbers in the relevant ISAT files that are not coded correctly. For all our analyses of ITBS scores in the 1990s, our retention rates for both the prereform and postreform cohorts are always between 80% and 82%. Because CPS administered these exams as part of their own accountability system, there were fewer problems with matching exams to correct student ID numbers. In the end, our ISAT analyses contain 18,305 and 19,651 students from the 2001 and 2002 samples, respectively, who have valid scores on both exams and were tested without accommodations in fifth grade. The rates of follow-up testing without accommodations are 0.673 for the 1999–2001 cohort and 0.706 for the 2000–2002 cohort.

Rates of follow-up testing increase with baseline achievement for both cohorts. This gradient reflects in large part our decision to exclude students who were tested with accommodations in fifth grade. Since our data do not record whether students were allowed to take the third-grade tests with accommodations or what types of accommodations fifth graders received, we are concerned that the introduction of NCLB could have affected the types of accommodations offered in ways that we cannot measure. For completeness, we did conduct similar analyses including the samples of accommodated fifth graders. This approach yields larger samples of students in both 2001 and 2002 whose previous third-grade scores signal low-baseline achievement, and given this approach, the estimated treatment effects associated with the bottom two deciles of baseline achievement are uniformly below those in figures 1A and 1B but within the confidence intervals presented.

Panel A of table A1 describes the data used to construct figures 1A and 1B and also describes differences in baseline characteristics between the 2001 and 2002 samples. The predicted fifth-grade scores for 2002 are based on the third-grade scores for the 2002 cohort and the estimated coefficients from regressions of fifth-grade math and reading scores on

TABLE A1.—TREATMENT AND CONTROL SCORES AND SAMPLE SIZES

A: 2002 vs. 2001 Samples, 5th-Grade ISAT

Decile	Sample Size		Follow-up Testing Rate		Average Math Score			Average Reading Score		
	2001	2002	2001	2002	2001 (Actual)	2002 (Actual)	2002 (Predicted)	2001 (Actual)	2002 (Actual)	2002 (Predicted)
3rd-grade score index (1999 sample)										
1	1,833	2,447	0.470	0.558	140.8	140.3	140.8	139.6	139.7	139.9
2	1,845	2,540	0.608	0.655	144.7	144.4	144.6	144.0	144.1	144.0
3	1,825	2,287	0.644	0.699	147.0	148.1	146.9	146.4	147.1	146.5
4	1,825	1,783	0.690	0.697	149.7	151.1	149.5	149.1	150.3	149.2
5	1,826	1,745	0.686	0.715	152.6	153.4	152.5	151.6	152.6	151.8
6	1,828	1,691	0.725	0.748	154.9	156.9	154.8	154.0	155.5	154.1
7	1,838	1,718	0.717	0.762	158.6	160.1	158.5	157.1	158.7	157.4
8	1,825	1,736	0.758	0.778	162.3	163.9	162.0	160.6	162.1	161.0
9	1,840	1,810	0.770	0.806	168.0	168.9	167.7	165.3	166.1	165.5
10	1,820	1,894	0.809	0.817	178.7	179.5	178.7	174.2	174.4	174.4
Total	18,305	19,651	0.673	0.706	155.7	155.5	154.6	154.2	154.0	153.4

B: 1998 vs. 1996 Samples, 5th-Grade ITBS

Decile	Sample Size		Follow-up Testing Rate		Average Math Score			Average Reading Score		
	1996	1998	1996	1998	1996 (Actual)	1998 (Actual)	1998 (Predicted)	1996 (Actual)	1998 (Actual)	1998 (Predicted)
3rd-grade score index (1994 sample)										
1	2,193	1,964	0.737	0.670	3.71	3.69	3.71	3.53	3.41	3.52
2	2,211	1,892	0.787	0.719	4.09	4.17	4.11	3.87	3.91	3.87
3	2,167	1,895	0.815	0.778	4.38	4.49	4.40	4.13	4.20	4.14
4	2,162	1,917	0.831	0.805	4.68	4.82	4.70	4.50	4.59	4.49
5	2,177	2,035	0.848	0.825	4.95	5.09	4.97	4.78	4.90	4.79
6	2,206	2,064	0.837	0.829	5.22	5.33	5.23	5.13	5.19	5.11
7	2,176	2,220	0.841	0.849	5.55	5.62	5.57	5.43	5.53	5.42
8	2,181	2,264	0.847	0.865	5.82	5.94	5.86	5.82	5.84	5.77
9	2,172	2,180	0.846	0.864	6.23	6.32	6.25	6.31	6.35	6.29
10	2,176	2,313	0.837	0.854	6.92	6.96	6.97	7.38	7.34	7.42
Total	21,821	20,744	0.821	0.804	5.15	5.30	5.24	5.09	5.20	5.16

C: 1998 vs. 1996 Samples, 6th-Grade ITBS

Decile	Sample Size		Follow-up Testing Rate		Average Math Score			Average Reading Score		
	1996	1998	1996	1998	1996 (Actual)	1998 (Actual)	1998 (Predicted)	1996 (Actual)	1998 (Actual)	1998 (Predicted)
4th-grade score index (1994 sample)										
1	2,406	2,370	0.720	0.696	4.40	4.45	4.37	4.04	4.07	4.01
2	2,314	2,092	0.795	0.740	4.93	5.15	4.96	4.56	4.76	4.57
3	2,366	2,206	0.811	0.787	5.28	5.53	5.30	4.94	5.18	4.95
4	2,370	2,263	0.835	0.818	5.62	5.88	5.64	5.28	5.51	5.28
5	2,342	2,245	0.834	0.848	5.92	6.16	5.95	5.63	5.79	5.63
6	2,372	2,389	0.841	0.855	6.20	6.46	6.21	5.94	6.12	5.94
7	2,351	2,338	0.843	0.865	6.54	6.77	6.56	6.30	6.45	6.30
8	2,364	2,416	0.855	0.868	6.86	7.07	6.88	6.71	6.86	6.70
9	2,362	2,543	0.847	0.861	7.37	7.51	7.36	7.26	7.33	7.28
10	2,348	2,551	0.844	0.861	8.20	8.28	8.20	8.46	8.54	8.50
Total	23,595	23,413	0.820	0.817	6.13	6.37	6.19	5.91	6.12	5.97

polynomials in the third-grade math and reading scores for the 2001 cohort. Because the 2002 cohort has slightly lower overall third-grade scores, the average predicted scores in 2002 are below those for 2001 in math and reading. Panels B and C of table A1 are similar descriptions of the data used to construct figures 3A and 3B, and 4A and 4B, respectively. Although panel A shows that higher rates of follow-up testing in the bottom deciles of achievement in the post-NCLB cohort, panels B and C show lower rates of follow-up testing in these deciles for the postreform cohorts. Here, the overall predicted average scores among the 1998 cohorts are slightly higher than the corresponding average scores among the prereform cohorts.

Having noted these differences in testing rates and average predicted scores by cohort, we stress that in all three panels, the average scores of the prereform cohorts and the average predicted scores for the postreform cohorts match well within deciles. Further, as we note in table 1, we have conducted our analyses including extra controls for gender, race, and eligibility for free lunch, and our results are almost identical. In the 1990s, our postreform cohorts are slightly more prepared on average than the prereform cohorts, and the opposite is true in the NCLB years, but within our decile groups, our treatment and control groups match well in terms of academic preparation during the prereform periods.

Some may worry that because the first decile follow-up testing rate in 2001 is 16% lower than the corresponding rate observed in 2002 (0.470 versus 0.558), our negative estimated achievement gains for this decile in 2002 could be driven by an increase in follow-up testing among low-performing students following NCLB. However, table A1 shows no evidence that this is the case with regard to observed characteristics. The expected scores for first-decile students in the 2002 cohort are the same as or slightly greater than those of first-decile students in 2001. Further, one would have to assume an incredible amount of selection on unmeasured traits within this first decile in order to alter the basic pattern of results in figures 1A and 1B. Even if one assumes that the average treatment effect among the additional 16% tested in 2002 is as low as -2, which is comparable in absolute value to the largest of our positive estimated treatment effects, the implied average treatment effects in math and reading for the balance of the sample remain less than -0.28 in math and less than 0.13 in reading.

APPENDIX B

Educational Triage

Here we show that if some students are receiving extra help, students who receive no extra help must be in the extremes of the ability distribution.

Recall the notation from section II. The school's problem is described by

$$\min_{e_i} \Psi\left(\sum_i F(\bar{\imath} - \alpha_i - e_i)\right) + \sum_i c(e_i) \qquad \text{s.t. } e_i \geq 0 \ \forall \ i = 1, 2, \ldots, N, \tag{A1}$$

where $\Psi(\cdot)$ and $c(\cdot)$ are both increasing, strictly convex functions. $f(\cdot)$ is unimodal.

Proposition. *For any three ability levels, $\alpha_H > \alpha_M > \alpha_L$, any effort plan e^* that satisfies:*

$$e_H^* > 0$$

$$e_M^* = 0$$

$$e_L^* > 0$$

cannot be a solution to equation (A1).

Proof. *Case 1:* Assume that $\bar{\imath} - \alpha_L - e_L^* < \bar{\imath} - \alpha_M$, or $\alpha_L + e_L^* > \alpha_M$. Define an alternative plan $\hat{e}$:

$$\hat{e}_L = e_M^* + (\alpha_M - \alpha_L)$$

$$\hat{e}_M = e_L^* - (\alpha_M - \alpha_L)$$

$$\hat{e}_H = e_H^*.$$

The penalties are the same for both $\hat{e}$ and e^* plans, but the total cost is higher for the e^* plan.

Case 2: $f'(\bar{\imath} - \alpha_M) \geq 0$. This implies that $f(\bar{\imath} - \alpha_M) \geq f(\bar{\imath} - \alpha_H - e_H^*)$. Thus, there exists $\delta > 0$ such that we can form an alternative plan $\hat{e}$:

$$\hat{e}_L = e_L^*$$

$$\hat{e}_M = \delta$$

$$\hat{e}_H = e_H^* - \delta$$

The penalties of $\hat{e}$ are weakly less than e^* and the costs are lower.

Case 3: $\alpha_L + e_L^* \leq \alpha_M$ and $f'(\bar{\imath} - \alpha_M) < 0$. These conditions imply that $f(\bar{\imath} - \alpha_M) \geq f(\bar{\imath} - \alpha_L - e_L^*)$. Thus there exists $\delta > 0$ such that we can construct an alternative effort plan $\hat{e}$:

$$\hat{e}_L = e_L^* - \delta$$

$$\hat{e}_M = \delta$$

$$\hat{e}_H = e_H^*.$$

Both the penalties and costs are lower for the alternative plan $\hat{e}$.