NBER WORKING PAPER SERIES

ACCOUNTABILITY, INCENTIVES AND BEHAVIOR: THE IMPACT OF HIGH-STAKES TESTING IN THE CHICAGO PUBLIC SCHOOLS

Brian A. Jacob

Working Paper 8968
http://www.nber.org/papers/w8968

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2002

I would like to thank the Chicago Public Schools, the Illinois State Board of Education and the Consortium on Chicago School Research for providing the data used in this study. I am grateful to Peter Arcidiacono, Anthony Bryk, Susan Dynarski, Carolyn Hill, Robert LaLonde, Lars Lefgren, Steven Levitt, Helen Levy, Susan Mayer, Melissa Roderick, Robin Tepper and seminar participants at various institutions for helpful comments and suggestions. Jenny Huang provided excellent research assistance. Funding for this research was provided by the Spencer Foundation. All remaining errors are my own. The views expressed herein are those of the author and not necessarily those of the National Bureau of Economic Research.

Accountability, Incentives and Behavior:
The Impact of High-Stakes Testing in the Chicago Public Schools
Brian A. Jacob
NBER Working Paper No. 8968
June 2002
JEL No. I20, I28, J24

ABSTRACT

The recent federal education bill, No Child Left Behind, requires states to test students in grades three to eight each year, and to judge school performance on the basis of these test scores. While intended to maximize student learning, there is little empirical evidence about the effectiveness of such policies. This study examines the impact of an accountability policy implemented in the Chicago Public Schools in 1996-97. Using a panel of student-level, administrative data, I find that math and reading achievement increased sharply following the introduction of the accountability policy, in comparison to both prior achievement trends in the district and to changes experienced by other large, urban districts in the mid-west. I demonstrate that these gains were driven largely by increases in test-specific skills and student effort, and did not lead to comparable gains on a state-administered, low-stakes exam. I also find that teachers responded strategically to the incentives along a variety of dimensions—by increasing special education placements, preemptively retaining students and substituting away from low-stakes subjects like science and social studies.

Brian A. Jacob
John F. Kennedy School of Government
Harvard University
79 JFK Street
Cambridge, MA 02138
and NBER
Brian_Jacob@harvard.edu

1. Introduction

In January 2002, President Bush signed the *No Child Left Behind Act of 2001* (NCLB), ushering in a new era of educational accountability. The new federal legislation requires states to test students in grades three through eight and to use these exam results to judge the performance of schools. If a school fails to make adequate progress for several consecutive years, the district must allow children to attend another public school in the district and provide students with supplemental education services such as private tutoring. Persistently low-performing schools may be closed or reconstituted with new staff and curriculum (Robelen 2002).

NCLB strengthens a movement toward accountability in education that has been gathering momentum for nearly a decade. Statutes in 25 states now explicitly link student promotion or graduation to performance on state or district assessments. At the same time, 18 states reward teachers and administrators on the basis of exemplary student performance and 20 states sanction school staff on the basis of poor student performance (Quality Counts 2002).

These accountability policies dwarf all other education reforms in scope. Consider, for example, one of the most popular school reform initiatives in recent years—school choice. Of the nearly 53 million children attending elementary and secondary schools in the country, only 60,000 used vouchers to attend a private school and 580,000 others attended a charter school percent of all schoolchildren (Howell and Peterson 2002, CER 2002). Of the roughly 47 million students in public schools, only four million participated in any type of public school choice program, which includes inter-district choice, magnet schools and other types of intra-district choice (NCES 1997). On the other hand, the accountability program in Texas alone impacts approximately 3.6 million students while the policies in Chicago and New York City affect an additional 1.5 million students. As the mandates of NCLB are implemented, all of the 33.4

million elementary students in the nation will be attending schools subject to test-based accountability.[1]

While the primary intent of such accountability policies is to provide incentives to maximize student learning, poorly designed incentives can have perverse consequences. For example, Holmstrom and Milgrom (1991) show that high-powered incentives will lead agents to focus on the most easily observable aspects of a multi-dimensional task. Based on similar logic, testing critics have argued that current accountability policies will cause teachers to shift resources away from low-stakes subjects, neglect infra-marginal students and ignore critical aspects of learning that are not explicitly tested.

Despite its increasing popularity within education, there is little empirical evidence on test-based accountability (also referred to as high-stakes testing, abbreviated hereafter as HST). The majority of existing research focuses on mandatory high school graduation exams, which provide incentives for secondary students but have little direct impact on teachers or administrators. Recent evidence on school-based accountability programs is mixed, with some studies showing modest achievement gains but other showing little change in student performance. Moreover, most studies of school-based accountability do not utilize individual student data and thus cannot examine many outcomes of interest or investigate how effects vary across students.

Test-based accountability raises three fundamental questions about the ways in which students and teachers respond to performance incentives. The most fundamental question about HST is whether it increases student achievement. Insofar as test-based accountability raises student motivation, increases parent involvement and/or improves curriculum or pedagogy, one

[1] All national enrollment figures are taken from the 2001 Digest of Education Statistics (Digest 2001).

would expect HST to improve student performance. Unfortunately, accountability policies are often implemented in conjunction with a variety of other reforms, frequently without any pre-existing data on student performance, making it difficult to attribute the achievement changes to the accountability policy.

Even if a positive causal relationship between HST and student achievement can be established, it is important to understand what factors are driving the improvements in performance. Critics of test-based accountability often argue that its primary impact is to increase the time spent on test-preparation activities, thus improving test-specific skills at the expense of more general skills. Others argue that test score gains reflect student motivation on the day of the exam. Thus, one might want to examine whether test score gains reflect increases in general skills, test-specific skills, transitory student effort or some combination thereof.[2] If HST increased the general skill level, observed achievement gains should be reflected in other measures of student outcomes. On the other hand, to the extent the improvements are due to transitory student effort or increases in test-specific skills, one might not expect the test results to generalize.

Finally, in evaluating the effectiveness of HST, it is important to understand whether teachers and administrators respond strategically to the incentives provided by the accountability policy. Critics have worried about educator responses along a number of dimensions. For example, since low-ability students bring down the performance level of a school, the policy provides an incentive for teachers to find ways to exclude students from testing. By placing low performing students in special education programs, teachers are able to exempt them from most

[2] Achievement gains may also be due to increases in cheating on the part of students, teachers or administrators. While Jacob and Levitt (2002) found that instances of classroom cheating increased substantially following the

standard testing and reporting procedures. If special education programs are ineffective or inappropriate for these students, this may have detrimental long-term effects on the development of low-ability students.

This paper addresses these questions in the context of a test-based accountability policy that was implemented in Chicago Public Schools (ChiPS) in 1996-97.[3] The ChiPS is an excellent case study for several reasons. First, Chicago was the first large, urban school district to implement high-stakes testing. Because the accountability policy was introduced in 1996-97, one can track student outcomes for up to four years. Second, detailed student level data is available for all ChiPS students with unique student identification numbers that allow one to track individual students over time. Earlier studies have relied on imperfect matching algorithms. This unique data set allows one to not only examine a variety of different outcomes, but also to investigate the heterogeneity of effects across students. Third, the Chicago policy resembles the policies being implemented throughout the country, incorporating incentives for both students and teachers. Beginning in 1996, Chicago schools in which fewer than 15 percent of students met national norms in reading were placed on probation. If student performance did not improve in these schools, teachers and administrators were subject to reassignment or dismissal. At the same time, the ChiPS took steps to end "social promotion," the practice of passing students to the next grade regardless of their academic ability. Students in third, sixth and eighth grades were required to meet minimum standards in reading and mathematics in order to advance to the next grade.

---

introduction of high-stakes testing in Chicago, they estimate that cheating increases could only explain an extremely small part of the test score gains since 1996-97.

[3] In this analysis, I do not focus on the programs that accompanied the introduction of the accountability policy such as summer school or training for teachers in low-achieving schools. For an evaluation of these programs, see Jacob

I find considerable evidence of a causal relationship between HST and student achievement. Math and reading scores on the city-administered Iowa Test Basic Skills (ITBS) increased sharply following the introduction of the accountability policy. These gains were substantially larger than would have been predicted by prior achievement trends in Chicago, and were substantially larger than the achievement changes experienced by other urban districts in Illinois and in other large mid-western cities. Moreover, the pattern of achievement gains is consistent with the incentives provided by the policy, with low-achieving schools showing substantially larger gains than other schools.

It appears that these achievement gains were driven primarily by increases in test-specific skills and student effort. There was no comparable jump in student test scores on the state-administered Illinois Goals Assessment Program (IGAP) following the introduction of the policy. Moreover, an item-level analysis of the ITBS math gains indicates that students made the greatest improvements on questions involving computation and number concepts—skills heavily emphasized on the ITBS exam—but little if any improvement on questions testing skills such as estimation, data interpretation and multiple-step problem-solving. While students made roughly equivalent improvement in all skill areas on the reading exam, they showed the largest improvement on test questions at the end of the exam (conditional on item difficulty), consistent with an increase in student effort during the exam (i.e., what might be described as a test "stamina" effect).

Finally, I show that teachers responded strategically to the incentives along a variety of dimensions. Following the introduction of high-stakes testing, (i) there was a substantial

---

and Lefgren (2002a, 2002b). For an earlier analysis of the accountability policy in Chicago, see Roderick, Jacob and Bryk (2001).

increase in the proportion of students in special education and/or excluded from testing; (ii) retention rates increased substantially in grades not directly affected by the student promotion policy; and (iii) science and social studies scores increased at a significantly slower rate than math and reading scores.

These findings have several interesting implications. On the one hand, they provide strong empirical support for general incentive theories, including the multi-task theories of Holmstrom and Milgrom (1991). Moreover, the findings from Chicago belie the view espoused by many policy-makers that teachers and schools are impervious to change. On the other hand, it is less clear how to evaluate high-stakes testing in Chicago as a school reform strategy. Because the achievement gains are driven largely by increases in skills emphasized on the ITBS exam, an assessment of the policy depends largely on how one values these skills and how much one believes that there has been a decrease in other skills that are not assessed on standardized achievement exams. One must also consider the impact of changes in special education and retention rates, which will depend on how one views these programs.

The remainder of this paper is organized as follows. Section 2 reviews the existing literature on high-stakes testing and provides some background on the Chicago policy. Section 3 discusses the empirical strategy and Section 4 describes the data. Sections 5 to 7 present the main findings and Section 8 concludes.

2. Background

2.1. *Prior Research on High-Stakes Testing*

The bulk of existing research on high-stakes testing focuses on high school graduation exams. While several studies have found a positive association between student achievement

and such exams (Bishop 1998, Frederisksen 1994, Neill 1998, Winfield, 1990), studies with better controls for prior student achievement find no achievement effects (Jacob 2001). However these studies provided only limited insight into impact of school-based accountability because they focus exclusively on high school students and do not involve policies that hold teachers or administrator accountable for student performance.

The evidence on school-based accountability programs and student performance is decidedly mixed. Craig and Sheu (1992) found modest improvements in student achievement after the implementation of a school-based accountability policy in South Carolina in 1984, but Ladd (1999) found that a school-based accountability program in Dallas during the early 1990s had few achievement benefits. Smith and Mickelson (2000) found that a similar program in Charlotte-Mecklenburg did not increase the academic performance of students relative to the state average. Several studies note that Texas students have made substantial achievement gains since the implementation of that state's accountability program (Grissmer and Flanagan 1998, Grissmer et. al. 2000, Haney 2000, Klein et. al. 2000, Toenjes et. al. 2000, Deere and Strayer 2001).

There is somewhat more consistent evidence that educators respond strategically to test-based accountability. Figlio and Getzler (2002) and Cullen and Reback (2002) find that schools respond to accountability policies by classifying more students as special needs or LEP (limited English proficient), thereby removing them from the test-taking pool. Koretz and Barron (1998) find survey evidence that elementary teachers in Kentucky shifted the amount of time devoted to math and science across grades to correspond with the subjects tested in each grade. Deere and Strayer (2001) found evidence that Texas schools have substituted across outputs in the face of

the TAAS system, focusing on the high-stakes subjects and low-achieving students.[4] Various studies suggest that test preparation associated with high-stakes testing may artificially inflate achievement, producing gains that are not generalizable to other exams (Linn and Graue 1990, Shepard 1990, Koretz et. al. 1991, Koretz and Barron 1998, Stecher and Barron 1998, Klein et. al. 2000).

2.2 High-Stakes Testing in Chicago

In 1996 the ChiPS introduced a comprehensive accountability policy designed to raise academic achievement. The first component of the policy focused on holding students accountable for learning, by ending a practice commonly known as "social promotion" whereby students are advanced to the next grade regardless of ability or achievement level. Under the new policy, students in third, sixth and eighth grades are required to meet minimum standards in reading and mathematics on the Iowa Test of Basic Skills (ITBS) in order to advance to the next grade.[5] Students who do not meet the standard are required to attend a six-week summer school program, after which they retake the exams. Those who pass move on to the next grade. Students who again fail to meet the standard are required to repeat the grade, with the exception of 15-year-olds who attend newly created "transition" centers.

The scope of the effort was one of the most striking features of Chicago's social promotion policy. Although many Chicago students in special education or bilingual programs are exempt from standardized testing, 70 to 80 percent of the students in the system were directly

[4] Deere and Strayer (2001) focus on TAAS gains, though Grissmer and Flanagan (1998) make a similar point regarding NAEP gains.

[5]The social promotion policy was actually introduced in Spring 1996 for eighth grade students, although it is not clear how far in advance students and teachers knew about this policy. In general, the results presented here remain

affected by the accountability policies. Of those who were subject to the policy, nearly 50 percent of third graders and roughly one-third of sixth and eighth graders failed to meet the promotional criteria and were required to attend summer school in 1997. Of those who failed to meet the promotional criteria in May, however, approximately two-thirds passed in August. As a result, roughly 20 percent of third grade students and 10 to 15 percent of sixth and eighth grade students were ultimately held back in the Fall.

In conjunction with the social promotion policy, the ChiPS also instituted a policy designed to hold teachers and schools accountable for student achievement. Under this policy, schools in which fewer than 15 percent of students scored at or above national norms on the ITBS reading exam were placed on probation. If they did not exhibit sufficient improvement, these schools could be reconstituted, which involved the dismissal or reassignment of teachers and school administrators. In 1996-97, 71 elementary schools serving over 45,000 students were placed on academic probation.[6] While ChiPS has only recently closed any elementary schools, teachers and administrators in probation schools as early as 1997 reported being extremely worried about their job security and staff in other schools reported a strong desire to avoid probation (Tepper, Stone & Roderick, forthcoming).

---

the same whether one considers the eighth grade policy to have been implemented in 1996 or 1997. Thus for simplicity, I use 1997 as the starting point for all grades.

[6] Probation schools received some additional resources and were more closely monitored by ChiPS staff. Jacob and Lefgren (2002b) examined the resource effects of probation using a regression discontinuity design that compared the performance of students in schools that just made the probation cutoff with those that just missed the cutoff. They found that the additional resources and monitoring provided by probation had no impact on math or reading achievement.

3. Empirical strategy

Because Chicago instituted its accountability policy district-wide in 1996-97, it is difficult to identify the causal impact of the program with certainty. Consider the following standard education production function:

$$(1) \qquad y_{isdt} = (HighStakes)_{dt}\,\delta + X_{isdt}\beta_1 + Z_{sdt}\beta_2 + u_s + \gamma_t + \eta_d + \phi_{dt} + \varepsilon_{isdt}$$

where y is an achievement score for individual i in school s in district d at time t, X is a vector of student characteristics, Z is a vector of school and district characteristics and ε is a stochastic error term. Unobservable factors are captured by student (u), time (γ), district (η) and time*district (ϕ) effects.

There are three primary threats to identification of δ, the effect of HST. First, one might be worried that the composition of students has changed substantially during the period in which HST was implemented, so that $Cov(HighStakes, u) \neq 0$. An influx of recent immigrants during the mid-to-late 1990s, for example, might bias δ downward whereas the return of middle-class students to the ChiPS would likely bias δ upward. Second, one might be concerned about changes at the state or national level that occurred at the same time as HST, so that $Cov(HighStakes, \gamma) \neq 0$. For example, state or federal education policies to reduce class size or mandate higher quality teachers that were enacted during the mid-1990s would likely lead us to overestimate the impact of HST. Similarly, improvements in the economy or other time-varying factors coincident with the policy would bias our estimates. Finally, one might be worried about other policies or programs in Chicago whose impact was felt at the same time as HST, so that

$Cov(HighStakes, \phi) \neq 0$. This includes programs implemented at the same time as HST as well as programs implemented earlier whose effects become apparent at the same time as the accountability policy was instituted (e.g., an increase in full-day kindergarten that began during the early 1990s).

The rich set of longitudinal, student-level data I use allows me to overcome some of these concerns. Using detailed administrative data for each student, I am able to control for observable changes in student composition, including race, socio-economic status and prior achievement. Moreover, because achievement data is available back to 1990, six years prior to the introduction of HST, I am also able to account for pre-existing achievement trends within the ChiPS. I thus look for a sharp increase in achievement (a break in trend) following the introduction of HST as evidence of a policy effect. Using data on students before and after the policy change, I estimate variations of the following specification:

$$(2) \qquad y_{ist} = (\text{HighStakes})\delta + (\text{PriorTrend})\gamma + X_{ist}\beta_1 + Z_{st}\beta_2 + \varepsilon_{ist}$$

This short, interrupted time-series design (Ashenfelter 1978) accounts for changes in observable characteristics as well as any unobservable changes (due to shifts in student composition, prior reform efforts in Chicago, and state or federal initiatives) that would have influenced student achievement in a gradual, continuous manner.[7] This is essentially a difference-in-difference estimator where the first difference is a within student change over time and the second difference is a district-wide change from pre-policy to post-policy. The size and

[7] The inclusion of a linear trend implicitly assumes that any previous reforms or changes would have continued with the same marginal effectiveness in the future. If this assumption is not true, the estimates may be biased. In addition, this aggregate trend assumes that there are no school-level composition changes in Chicago. I test this assumption by including school-specific fixed effects and school-specific trends in certain specifications and find comparable results.

scope of the accountability policy in Chicago mitigates any concern about other district-wide programs that might have been implemented at the same time as HST.[8]

One drawback of this strategy is that it does not account for time-varying effects that would have influenced student achievement in a sharp or discontinuous manner. One might be particularly concerned about unobservable changes on the state or national level effecting student performance (e.g., implementation of state or federal school reform legislation). Also, if there is substantial heterogeneity in the responses to the policy, then the achievement changes may appear more gradual and be harder to differentiate from other trends in the system. For example, if certain schools believed that the policy was temporary and therefore did not substantially change their behavior during the first year of the policy.

I attempt to address these concerns using a panel of achievement data on other urban districts in Illinois (e.g., Springfield, Peoria) as well as large mid-western cities outside of Illinois (e.g., St. Louis, Milwaukee, Cincinnati). I estimate variations of the following specification:

$$\bar{y}_{dt} = (HighStakes)_{dt}\,\delta + \Gamma X_{dt} + \Pi Z_{dt} + \varepsilon_{dt} \tag{3}$$

where $\bar{y}$ is the average reading or math score for district *d* at time *t, HighStakes* indicates the presence of high-stakes testing, X is a vector of district-specific fixed effects and district-specific trends, and *Z* is a vector of time-varying district characteristics (including aggregate student characteristics). This too is essentially a difference-in-difference estimator where the first difference is the district-level change from pre-policy to post-policy and the second difference is a comparison of changes across districts.

[8] While there were smaller programs introduced in Chicago after 1996, these were generally part (or a direct result) of the accountability policy. I simply assume that the effects of these policies are part of the HST impact.

4. Data

This study utilizes detailed administrative data from the ChiPS. Student records include information on a student's school, home address, demographic and family background characteristics, special education and bilingual placement, free lunch status, standardized test scores, grade retention and summer school attendance. More importantly, student identification numbers allow one to follow students across years as long as they remain in the ChiPS, so that I do not have to rely on imperfect matching strategies.[9] ChiPS personnel and budget files provide information on the financial resources and teacher characteristics in each school and school files provide aggregate information on the school population, including daily attendance rates, student mobility rates and racial and SES composition.

The measure of achievement used in Chicago is the Iowa Test of Basic Skills (ITBS), a standardized, multiple-choice exam developed and published by the Riverside Company. Student scores are reported in grade equivalents that reflect the years and months of learning a student has mastered. The exam is nationally normed so that a student at the 50th percentile in the nation scores at the eighth month of her current grade (e.g., an average third grader will score a 3.8). In order to compare achievement gains across grade level and to provide a way to interpret the magnitude of Chicago gains, I standardize all achievement scores separately by grade using the 1993 student-level mean and standard deviation.

The primary sample used in this analysis consists of students who were in 3rd, 6th and 8th grade from 1993 to 2000. For most analyses, I limit the sample to first-time students (e.g., students in the third, sixth or eighth grade for the first time in their school career) because the implementation of the social promotion policy caused a large number of low-performing students

in third, sixth and eighth grade to be retained, which substantially changed the student composition in these and subsequent grades beginning in 1997-98.[10] In order to have sufficient prior achievement data for all students, I limit the analysis to cohorts beginning in 1993.

I delete less than 2 percent of the students from the sample because they were missing demographic information. In addition, each year roughly 10 percent of students were not tested (most often because of a special education or bilingual placement) and are therefore not included in the achievement estimates, although they are included in the estimates of other outcomes.[11] To avoid dropping students with missing prior achievement data, I impute prior achievement using other observable student characteristics and create a variable indicating that the achievement data for that student was imputed.

Table 1 presents summary statistics for the sample. Like many urban school districts across the country, Chicago has a large population of minority and low-income students. In the sample of third, sixth and eighth graders from 1993 to 1996, for example, roughly 55 percent of students are Black, 30 percent are Hispanic and nearly 80 percent receive free or reduced price lunch. During this period, roughly 12 percent of all students were in special education programs and 13 percent were either not tested or had scores that were not included for official reporting purposes (generally because of a bilingual or special education placement). Among students who were tested, Chicago students scored roughly three-quarters of a year below national norms in math and nearly one year below national norms in reading. Looking across columns, we see that

---

[9] There is no significant change in the percent of students leaving the ChiPS (to move to other districts, to transfer to private schools, or to drop out of school) following the introduction of the accountability policy.

[10] While focusing on first-timers allows a consistent comparison across time, it is still possible that the composition changes generated by the social promotion policy could have affected the performance of students in later cohorts. For example, if first-timers in the 1998 and 1999 cohorts were in classes with a large number of low-achieving students who had been retained in the previous year, they might perform lower than otherwise expected. This would bias the estimates downward.

there were some changes in the student composition during the 1990s. There were slight increases in the percentage of Hispanic students in the ChiPS and increases in the percent of students living in foster care, participating in bilingual programs and receiving free or reduced price lunch. On the other hand, there was some increase in initial student achievement—e.g., prior reading achievement increased from an average of 0.89 grade equivalents below norms to 0.71 grade equivalents below norms. Perhaps more importantly, there were dramatic increases in math and reading achievement under high-stakes testing, with students gaining roughly 0.50 GE's in math and 0.40 GE's in reading. However, special education rates also increased, from 0.116 to 0.139.

5. Did high-stakes testing increase student achievement in Chicago?

5.1 Math and Reading Trends on the ITBS

Figure 1 shows unadjusted math and reading achievement trends in Chicago from 1990 to 2000, combining the data from grades three, six and eight and standardizing student test scores using the 1990 student-level mean and standard deviation. Following a slight decline in the early 1990s, test scores increased in 1993 and remained relatively constant until 1995 or 1996, after which they began to increase. The jump in 1993 is likely due to a new form of the ITBS introduced that year. The ChiPS administered several different forms of the ITBS throughout the 1990s, rotating the forms so that identical forms were never administered in two consecutive

[11] Among those students who are included in the achievement analysis less than two percent are deleted for missing demographic data.

years. In fact, this is the primary reason for some of the year-to-year choppiness in the trends.[12] Some teachers report that the form of the reading exam administered in 1997—the first year of the accountability policy—was more difficult than earlier exams, which may explain why observed reading scores did not increase substantially in 1997. (To obtain a cleaner picture of changes in student performance, one can compare cohorts taking identical forms – 1994, 1996 and 1998 or 1993, 1995 and 2000. As I show later, the general findings of the analysis remain the same if one focuses on achievement changes across identical test forms.)

The raw test score trends suggest that achievement began to increase somewhat prior to the introduction of the accountability policy. One explanation for this is that educators may have made changes in anticipation of the new policy. When the new superintendent assumed responsibility of the ChiPS in 1996, he made it clear that his administration would focus on improving achievement and would be holding schools accountable for student performance on standardized tests. While this finding is consistent with an anticipation effect, it is also possible that the early improvements in student achievement resulted from changes in student composition or earlier reform efforts.

To control for changes in student composition and prior achievement levels, Figure 2 plots the predicted versus observed achievement scores for successive cohorts of Chicago students from 1993 to 2000.[13] The predicted values are derived from an OLS regression model that includes cohorts 1993 to 1996 and controls for student, school and neighborhood demographics as well as prior academic achievement and a linear time trend. The trends in

[12] Different forms of the exam are supposedly equated, but teachers and administrators acknowledge that forms still vary slightly in difficulty.

[13] Test scores are standardized separately by grade using the 1993 mean and student standard deviation. The earliest years are excluded from the series because it is not possible to obtain prior achievement measures for students in these cohorts.

Figure 2 suggest that neither observable changes in student composition nor pre-existing trends in Chicago can explain the substantial improvement in student performance since 1997. In math, we see that observed achievement seemed to decrease somewhat from 1993 to 1996, but then increased sharply after 1996. In contrast, predicted achievement decreases slightly or remains flat over this period. By 2000, observed math scores are roughly 0.30 standard deviations higher than predicted. A similar pattern is apparent in reading. Predicted and observed test scores are relatively flat from 1993 to 1996. In 1997, the gap between observed and predicted scores appears to widen somewhat and grows substantially in 1998. By 2000, students are scoring roughly 0.20 standard deviations higher than predicted.

Even if there were no appreciable change in student composition in Chicago, it could be that the achievement gains in Chicago reflect more general improvements in student performance in the state or nation. The economy was improving throughout the later half of the 1990s, and there was a considerable emphasis on public education at the federal level. Student achievement nationwide, as measured by the National Assessment of Educational Progress (NAEP), increased roughly 0.25 standard deviations in math during the 1990s, although there was no gain in reading.[14]

To control for unobserved, time-varying factors at the state and/or national level, Figure 3 shows the Chicago trends relative to other urban school districts in Illinois and to other large, mid-western cities including Cincinnati, Gary, Indianapolis, Milwaukee and St. Louis, none of which implemented a comparable accountability policy during this period. The district-level averages are standardized using the student-level mean and standard deviation from the earliest possible year for each grade*subject*district (most often 1993). The Chicago and comparison

group trends track each other remarkably well from 1993 to 1996, and then begin to diverge in 1997. Math and reading achievement in the comparison districts fluctuates somewhat, but remains relatively constant from 1996 to 2000. In contrast, the achievement levels in Chicago rise sharply over this period.

Together, these figures suggest that the accountability policy in Chicago led to a substantial increase in math and reading achievement. To provide a more precise estimate of the effects, Table 2 shows the OLS regression results that correspond to Figures 2. Control variables include race, gender, race*gender interactions, guardian, bilingual status, special education placement, prior math and reading achievement, school demographics (including enrollment, racial composition, percent free lunch, percent with limited English proficiency and mobility rate) and demographic characteristics of the student's home census tract (including median household income, crime rate, percent of residents who own their own homes, percent of female-headed household, mean education level, unemployment rate, percent below poverty, percent managers or professionals and percent who are living in the same house for five years). Prior achievement is measured by math and reading scores three years prior to the base year (i.e., at *t-3*). This is done to ensure that the prior achievement measures are not endogenous. Because the 1999 cohort of sixth graders experienced high-stakes testing beginning in 1997, for example, one would not want to include their fourth or fifth grade scores in the estimation.[15] I include second and third order polynomials in prior achievement in order to account for any non-linear relationship between past and current test scores.

---

[14] Author's calculation based on data available from the National Center of Education Statistics (www.nces.ed.gov).
[15] For the 2000 cohort, test scores at *t-3* are endogenous as well. As a practical matter, however, it does not appear to make any difference whether one uses prior achievement at *t-3* or *t-4*, so I have used *t-3* in order to include as many cohorts as possible.

The estimates in Table 2 reveal several interesting findings. First, the policy effect appears to increase from 1997 to 2000. This is consistent with the fact that the later cohorts experienced more of the "treatment" and with the possibility that students and teachers may have become more efficient at responding to the policy over time. It is not possible to distinguish between these hypotheses because the policy was implemented district-wide in 1996-97. Second, it appears that the effects are somewhat larger for math than reading. This is consistent with a number of education evaluations that show larger effects in math than reading, presumably because reading achievement is determined by a host of family and other non-school factors while math achievement is determined largely by school. Third, it appears that the effects are somewhat larger for 8th grade students. This is consistent with the fact that eighth graders faced the largest incentives (they cannot move to high school with their peers if they fail to meet the promotional standards) and that they may be most able to influence their own learning.[16] Table 3 shows the estimates reflecting the comparison between Chicago and the comparison districts. These results suggest that the accountability policy in Chicago increased student math achievement by roughly 0.35 standard deviations and reading achievement by 0.25 standard deviations. [17]

[16] This result must be interpreted with caution since some observers have questioned whether the grade equivalent metric can be compared across grades (Petersen et. al. 1989; Hoover 1984). Roderick et. al. (2001) attempt to correct for this and find similar results.

[17] One additional factor is important to note in interpreting these results. The estimates for the latter cohorts may be biased because of compositional changes resulting from grade retention. For example, the 1999 and 2000 eighth grade cohorts will not include any students who were retained as sixth graders in 1997 or 1998. To the extent that retention is correlated with unobservable student characteristics that directly affect achievement, this will bias the estimates. However, Jacob and Lefgren (2002a) found little difference between OLS and IV estimates of summer school and grade retention, suggesting that there may *not* be much significant correlation (conditional on prior achievement and other observable characteristics). However, even if they were not retained, a proportion of the students in these cohorts will have attended summer school as sixth graders, which Jacob and Lefgren (2002a) show to increase subsequent achievement. Therefore, it is best to interpret these coefficients for the later cohorts as upper bounds on the incentive effect of the policy.

5.2 The Heterogeneity of Effects Across Student and School Risk Level

If the improvements in student achievement were caused by the accountability policy, one might expect them to vary across students and schools. In particular, one might expect marginal students and schools to show the largest achievement gains since the policy will be binding for them and they will likely feel that they have a reasonable chance of meeting the standard. Three margins are relevant: (1) the social promotion margin—in order to be promoted, students were required to achieve at roughly the 20^{th} percentile (on the national ability distribution) in reading and math; (2) the student margin for probation—a school's probation status is determined by the percent of students that score above the 50^{th} percentile nationally in reading; and (3) the school probation margin—in order to avoid probation, 15 percent of students in the school must meet national norms in reading.

In order for teachers and administrators to translate these incentives into differential achievement effects, several conditions must hold. First, production must be divisible. That is, schools must be able to focus attention on certain students and not others, perhaps by providing individualized instruction. If schools rely on class- or school-wide initiatives such as curriculum changes, test preparation or student motivation, then they may not be able to effectively target specific students. Second, the main effect of teacher or student effort must be large relative to that of initial ability or the interaction between effort and initial ability. If teacher effort has a substantially larger effect on high ability students than low ability students, then HST may result in larger gains for higher ability students despite the structure of the incentives. Finally, schools must be able to clearly distinguish between high and low ability students. While this may seem trivial given the prevalence of achievement testing in schools, sampling variation and

measurement error in achievement exams may expand the group of students viewed as "marginal" by teachers and students.

To examine the changes in achievement across student abilities, Table 4 shows OLS estimates of the differential effects across students and schools. Prior student achievement is based on the average math and reading score three years prior to the baseline test year (i.e., 5th grade scores for the 8th grade cohorts).[18] Prior school achievement is based on the percent of students in the school in 1995 that met national norms on the reading exam.[19] The sample includes first-time students whose scores were included for reporting purposes. The latest cohorts are excluded from the sample because these students will have experienced previous retentions, which may bias the results. The regressions also include the full set of control variables used in Table 2.

Model 1 provides the average effect for all students in all of the post-policy cohorts, providing a baseline from which to compare the other results. Model 2 shows how the effects vary across student and school risk level. Note that the omitted category includes the highest ability students (those who scored above the 50th percentile in prior years) in the highest achieving schools (schools where at least 40% of students were meeting national norms in prior years). Looking across all grades and subjects, several broad patterns become apparent. First, students in low-performing schools seem to have fared considerably better under the policy than comparable peers in higher-performing schools. In sixth grade math, for example, students in the schools where fewer than 20 percent of students had been meeting national norms in previous years gained 0.159 standard deviations more than comparable peers in schools where over 40

[18] Second grade test scores are used to determine prior achievement for third graders since this is the first year that the majority of students take the standardized achievement exams.
[19] The results are robust to classifying school risk on the basis of achievement in other pre-policy years.

percent of students had been meeting national norms. This makes sense since the accountability policy imposed much greater incentives on low-performing schools that were at a real risk of probation.

Second, students who had been scoring at the 10^{th}-50^{th} percentile in the past fared better than their classmates who had either scored below the 10^{th} percentile, or above the 50^{th} percentile. This is consistent with the incentives imposed on at-risk students by the policy to end social promotion. Moreover, the effect for marginal students appears somewhat stronger in reading than math, suggesting that there may be more intentional targeting of individual students in reading than in math, or that there is greater divisibility in the production of reading achievement. However, it is also important to note that the differential effects of student prior ability are considerably smaller than the differential effects of prior school achievement. This suggests that responses to the accountability policy took place at the school level, rather than the individual student level.[20]

5.3 Student-Focused versus School-Focused Accountability

Unlike most previous accountability systems, high-stakes testing in Chicago provided direct incentives for students as well as teachers. Students in third, sixth and eighth grade were required to pass reading and math exams to move to the next grade, while schools were judged on the basis of the reading performance of students in grades three to eight. By examining differential gains across subject and grade, one might theoretically separate the effect of the student and school-based accountability policy. However, in practice there are several

[20] This result may also be due to measurement error, although this seems somewhat less likely because the student prior achievement measure is an average of two exam scores—math and reading—and similar results were obtained using a measure composed of several earlier years of test data.

difficulties in this type of comparison. First, because the lowest-achieving third and sixth graders were retained beginning in 1997, the subsequent cohorts in grades four, five and seven will be composed of substantially higher-achieving students. Second, many of the 1998 fourth and seventh graders attended summer school the previous year exposing them to two additional months of instruction.

Given these concerns, the 1997 cohort will provide the most easily interpretable comparison between gate and non-gate grades. Table 5 presents the policy affects for grades three, six and eight (i.e., promotional gate grades) versus grades four, five and seven (i.e., non-gate grades). In 1997, there appears to be little difference in achievement effects between students in the promotional gate grades and those not subject to the promotional gate in 1997.[21] One explanation for this finding is that the school probation policy was driving the overall achievement results. An alternative explanation is that students in grades four, five and seven incorrectly believed that they were subject to the promotional requirements. Student interviews provide some evidence for such confusion, possibly because teachers in these grades used accountability as a classroom management tool, emphasizing the promotional criteria to motivate students.[22] A third explanation rests on indivisibilities in production within elementary schools. For example, restructuring the school day to allow more time for math and reading may necessarily involve all grades in the school. Finally, it is possible that the first year effects were somewhat anomalous, perhaps because students and teachers were still adjusting to the policy or because the form change that year may have affected grades differentially. Because it was not affected by composition changes, grade five provides a reasonable comparison for the gate grades

[21] The results are similar across the ability distribution. Tables available from the author upon request.
[22] For more information on qualitative studies of the accountability policy in Chicago, see Engel and Roderick (2001).

in 1998. The 1998 accountability effects are at least twice as large in grades three, six and eight compared with grade five (for example, 0.144 versus 0.067 s.d. gain in math), suggesting that the student accountability provisions may have played a large role in the overall policy in later years.

6. What factors are driving the improvements in performance in Chicago?

Even if a positive causal relationship between HST and student achievement can be established, it is important to understand what factors are driving the improvements in performance. Critics of test-based accountability often argue that the primary impact of HST is to increase the time spent on test-specific preparation activities, which could improve test-specific skills at the expense of more general skills. Others argue that test score gains reflect student motivation on the day of the exam. Unfortunately, because such things as effort and test preparation are not directly observable, it is difficult to disentangle the factors underlying the achievement gains in Chicago. This section attempts to shed some light on the factors driving the achievement gains in Chicago, first by comparing student performance across exams and then by examining the ITBS improvements in greater detail.

6.1 The Role of General Skills

Even the most comprehensive achievement exam can only cover a fraction of the possible skills and topics within a particular domain. Because all standardized tests differ to some extent in format and content, one would not expect gains on one test to be completely reflected in performance changes on another exam. Differences in student effort across exams (or rather changes in student effort) also complicate the comparison of performance trends from one test to another. Nonetheless, it is instructive to compare achievement changes on the high-stakes exam

to changes on alternate tests since this will provide information on the extent to which improvements in general versus test-specific skills were driving the observed test score gains.

Under the Chicago accountability policy, student promotion and school probation are based entirely on the Iowa Test of Basic Skills (ITBS), an exam that has been administered by the district for many years. Chicago students also take a state-administered achievement exam known as the Illinois Goals Assessment Program (IGAP). While the two exams have a similar format (they are both timed, multiple-choice exams), the IGAP reportedly places somewhat greater emphasis on critical thinking and problem-solving skills.[23]

If the accountability policy operated by increasing general skills, or a broad enough range of specific skills, the observed ITBS gains in Chicago should be reflected to some extent in the IGAP trends. Figure 4 shows IGAP achievement trends in Chicago relative to other urban districts in Illinois.[24] The data for this analysis is drawn from school "report cards" compiled by the Illinois State Board of Education (ISBE) which provide average IGAP scores by grade and subject as well as background information on schools and districts.[25] The analysis is limited to the period from 1993 to 1998 because Illinois introduced a new exam in 1999. The Chicago sample excludes students retained under the new promotional policy in order to provide a valid

[23] The IGAP math exam has fewer straight computation questions, and even these questions are asked in the context of a sentence or word problem. Similarly, with long passages, multiple correct answers and questions asking students to compare passages, the IGAP reading exam appears to be more difficult and more heavily weighted toward critical thinking skills than the ITBS exam.

[24] To identify the comparison districts, I first identify districts in the top decile in terms of the percent of students receiving free or reduced price lunch, percent minority students, and total enrollment and in the bottom decile in terms of average student achievement (averaged over third, sixth and eighth grade reading and math scores) based on 1990 data. Not surprisingly, Chicago falls in the bottom of all four categories. Of the 840 elementary districts in 1990, Chicago ranks first in terms of enrollment, 12th in terms of percent of low-income and minority students and 830th in student achievement. Other districts that appear at the bottom of all categories include East St. Louis, Chicago Heights, East Chicago Heights, Calumet, Joliet, Peoria and Arora. I then use the 34 districts (excluding Chicago) that fall into the bottom decile in at least three out of four of the categories. I have experimented with several different inclusion criteria and the results are not sensitive to the choice of the urban comparison group.

comparison with other districts. The achievement measure is standardized using the school level mean and standard deviation in Illinois in 1993.

In 1993, Chicago students scored between 0.40 and 0.80 standard deviations below students in other urban districts, but appear to have narrowed the achievement gap during the mid-1990s.[26] However, at least in grades three and six, this trend appears to have begun prior to the introduction of high-stakes testing in these grades and there was no noticeable break in trend in 1997. Achievement scores in grade eight, particularly in reading, show some break beginning in 1996.[27] Table 6 shows corresponding OLS estimates that control for a variety of time-varying school and district characteristics including racial composition, percent of students receiving free or reduced price lunch, the percent of Limited English Proficient (LEP) students, school mobility rates, per-pupil expenditures in the district and the percent of teachers with at least a Masters degree in the district. The coefficient estimates shown in the table reflect the interaction between high-stakes testing years (1997 and 1998) and an indicator variable for Chicago. The point estimates indicate that once we take into account district-specific pre-existing trends and demographics, HST appears to have a slight negative effect on IGAP achievement in Chicago. Rows 4 and 5 that show estimates based on the Chicago schools alone tell a similar story.

As on the ITBS, low-achieving schools made larger gains on the IGAP than high-achieving schools. Table 7 shows estimates for grades three, six and eight by school

---

[25] The Illinois State Board of Education (ISBE) does not provide item-level achievement results for the IGAP so it is not possible to conduct a detailed analysis of IGAP improvement, or to directly compare ITBS and IGAP gains for similar questions.

[26] One explanation for this is that the IGAP was viewed as the high-stakes exam prior to 1995. The state publishes IGAP results annually and each year local newspapers run lengthy articles comparing results across schools and districts. After 1993, the Illinois State Board of Education (ISBE) began reporting student level IGAP scores to schools and parents for the first time, and in 1995 the ISBE began using IGAP results to place low-achieving schools on a state watch list. During this period, the ChiPS placed little if any emphasis on the ITBS.

[27] This is one case where it does appear important to recognize that the accountability policy started for eighth graders in 1996.

achievement level. In the first row, the sample includes only Chicago schools, which are divided into the same three categories used earlier (i.e., bottom schools are those in which 0-20% of students were meeting national reading norms on the ITBS in 1995, middle schools had 21-40% students meeting national norms, and top schools had more than 40% meeting norms). In the lowest-achieving schools, the IGAP scores showed no statistically significant change following the introduction of HST. In contrast, IGAP scores in the top schools dropped roughly 0.14 and 0.13 standard deviations in reading and math. The second row presents estimates using the urban comparison districts). Here the schools are grouped into three equal size groups on the basis of their aggregate IGAP scores in the early 1990s. While few of these estimates are statistically significant, the point estimates suggest a similar pattern, with lower-achieving schools doing relatively better on the IGAP under high-stakes testing.

6.2 The Role of Specific Skills

If the ITBS gains were not driven primarily by an increase in general skills, it is possible that they were the result of improvements in ITBS-specific skills. Based on analysis of teacher survey data, Tepper (2002) concluded that ITBS-specific test preparation and curriculum alignment increased following the introduction of the accountability policy. One way to examine the importance of these factors is to compare improvement across test items. To the extent that the disproportionately large ITBS gains were driven by ITBS-specific curriculum alignment or test preparation, we might expect to see the largest gains on ITBS items that are (a) easy to teach and/or (b) relatively more common on the ITBS than the IGAP. In math, these include questions that test computation and basic number concepts (e.g., arithmetic with negative and positive numbers, ordering numbers in sequence, using place value and scientific notation, etc.).

Table 8 presents OLS estimates of the relationship between high-stakes testing and ITBS math achievement by item type. The sample includes grades three, six and eight. By focusing on only those cohorts that took Form L, this analysis allows one to compare student performance on identical questions over time. The dependent variable is the proportion of students who answered the item correctly in the particular year. Note that these specifications also include controls for item difficulty to account for the correlation between item type, position and difficulty (e.g., the fact that the more difficult items are often included at the end of the exam and that certain types of questions are inherently more difficult for students).[28]

Column 1 classifies questions into two groups—those testing basic skills such as math computation and number concepts and those testing more complex skills such as estimation, data interpretation and problem-solving (i.e., word problems). Students in 1998 were 1.7 percentage points more likely to correctly answer questions involving complex skills in comparison to cohorts in 1994 and 1996. The comparable improvement for questions testing basic skills was 3.9 percentage points, suggesting that under accountability students improved more than twice as much in basic skills as compared with more complex skills. Column 2 separates items into five categories—computation, number concept, data interpretation, estimation and problem-solving—and shows the same pattern. In column 3, the items are classified into very detailed categories, providing even more information on the relative gains within the math exam. Student performance on items involving whole number computation (the omitted category) increased 3.5 percentage points. Interestingly, students improved even more, nearly 5.7 percentage points, on items involving computation with fractions. Questions testing knowledge of probability and

[28] The item difficulty measures are the percentage of students correctly answering the item in a nationally representative ample used by the test publisher to norm the exam. Interactions between item difficulty and the

statistics also appear to have made relatively large gains. In contrast, students appear to have made no improvement on questions involving estimating compensation (problems involving currency) and the effective use of various strategies to solve word-problems, and very little (if any) improvement on items involving multiple-step word problems, measurement and interpreting relationships shown in charts, graphs or tables.

Table 8 presents similar estimates for reading. The first column includes no indictor for item type while columns 2 and 3 include increasing more detailed item-type classifications. Unlike math, it appears that the improvements in reading performance were distributed equally across question type. This analysis suggests that test preparation may have played a large role in the math gains, but was perhaps less important in reading improvement. One reason may be that it is relatively easier to teach specific math skills whereas reading instruction in the elementary grades may focus largely on phonics, practice reading or other activities that are not specifically geared to particular test items. Another explanation is that reading skills are more likely than math skills to be learned out of school.

6.3 The Role of Effort

Student effort is another likely candidate for explaining the large ITBS gains. Interview and survey data provide evidence that students, particularly students in the sixth and eighth grades, were acutely aware of and worried about the accountability mandates (Tepper 2002; Roderick and Engel 2001; Tepper, Stone and Roderick, Forthcoming). If the consequences associated with ITBS performance led students to concentrate more during the exam or caused

---

accountability regime (1998 cohort) are included as well. The coefficients on the item difficulty*high-stakes interactions are generally insignificant.

teachers to ensure optimal testing conditions for the exam, test scores may have increased regardless of changes in general or test-specific skills.[29]

Test completion is one indicator of effort. Prior to the introduction of high-stakes testing, roughly 20 percent of students left items blank on the ITBS reading exam and nearly 38 percent left items blank on the math exam, despite the fact that there was no penalty for guessing.[30] If we believe that ITBS gains were due largely to guessing, we might expect the percent of questions answered to increase, but the percent of questions answered *correctly* (as a percent of all *answered* questions) to remain constant or perhaps even decline. However, from 1994 to 1998, the percent of questions answered increased by 1 to 1.5 percentage points while the percent correct as a fraction of the percent answered increased by 4 to 5 percentage points, suggesting that the higher completion rates were not due entirely to guessing. This pattern is true even among the lowest achieving students who left the greatest number of items blank prior to the accountability policy. Even if we were to assume that the increase in item completion is due entirely to random guessing, however, guessing could only explain 10 to 15 percent of the observed ITBS gains (Jacob 2002).

While increased guessing cannot explain a significant portion of the ITBS gains, other forms of effort may play a larger role. Insofar as there is a tendency for children to "give up" toward the end of the exam—either leaving items blank or filling in answers randomly—an increase in effort may lead to a disproportionate increase in performance on items at the *end* of the exam. One might describe this type of effort as test stamina—the ability to continue working and concentrating throughout the entire exam. In order to identify test stamina effects, the

[29] This might also be considered an effect of better testing conditions. Figlio and Winicki (2002) present evidence that schools attempt to enhance testing conditions by altering the content of meals served to students during testing.
[30] The math exam consists of three subsections and is thus roughly three times as long as the reading exam.

estimates in Tables 8 and 9 include variables indicating the item position—specifically, dummy variables denoting into which quintile of the exam the item falls (recall these estimates are conditional on item difficulty). In math, we see no relationship between item position and improvement under accountability. This is most likely because the math exam is divided into several sections so that each section is relatively short. In reading, on the other hand, this relationship is striking. Under the accountability policy, student performance on items at the end of the exam increased significantly more than performance on items at the beginning of the exam. In column 1, for example, we see that students in 1998 were 3.6 percentage points more likely to answer the first 20 percent of items on the exam, as compared with students in 1994 and 1996. Comparing the gain across item position groups, we see that 1998 students improved nearly 6.7 percentage points on the final 20 percent of items. Thus, student performance on the last 20 percent of items increased nearly twice as much as on the first 20 percent of items under the accountability policy. This effect remains the same as one includes increasingly detailed item type information in columns 2 and 3. This suggests that effort may have played a significant role in the ITBS gains seen under high-stakes testing.

6.4. Summary

The improvement in math achievement in Chicago appears to be driven largely by gains in specific skill areas such as math computation that make up a large portion of the ITBS, but are emphasized less on the IGAP. This suggests that teachers aligned their math curriculum to more closely match the content of the high-stake exam. In reading, ITBS gains were equally distributed across item types, but were considerably larger among questions at the end of the exam. This suggests that student effort or "stamina" played a larger role than test preparation in

the observed reading improvements. The fact that IGAP trends did not jump sharply following the introduction of the accountability policy confirms that the ITBS gains were not driven entirely by improvements in general skills. However, it is important to recognize that IGAP scores continued to increase in an absolute sense, which may mean that there was no substantial tradeoff in terms of skills.

7. Did educators respond strategically to high-stakes testing?

In evaluating the effectiveness of HST, it is important to understand whether teachers and administrators respond strategically to the incentives provided by the accountability policy. Critics of test-based accountability worry about educator responses along a number of dimensions, ranging from changes in the rate of special education placements to substitution away from low-stakes subjects. This section examines several of these issues.

7.1 Low-stakes versus high-stakes subjects

Given the consequences attached to test performance in certain subjects, one might expect teachers and students to shift resources and attention toward subjects included in the accountability program. We can test this theory by comparing trends in math and reading achievement after the introduction of HST with test score trends in social studies and science, subjects that are not included in the Chicago accountability policy. Unfortunately science and social studies exams are not given in every grade, and the grades in which these exams are given

has changed over time. For this reason, we limit the analysis to grades four and eight, from 1995 to 1998.[31]

Table 10 shows the impact of the accountability policy on a variety of subjects. Achievement gains in math and reading were roughly two to four times larger than gains in science and social studies, although science and social studies scores also increased under HST. The distribution of effects is also somewhat different for low versus high-stakes subjects. As we noted earlier, in math and reading, students in low-achieving schools experienced greater gains. , However, conditional on school achievement, low-ability students appeared to make only slightly larger gains than their peers. In science and social studies, on the other hand, low ability students showed significantly lower gains than their higher-achieving peers, while school achievement had little if any effect on science and social studies performance. This suggests that schools were shifting resources across subjects, particularly for low-achieving students, which is consistent with findings by Koretz and Barron (1998) and Deere and Strayer (2001).

7.2 Special education placements

While the accountability policies in Chicago are designed to increase student achievement, they also create incentives for teachers and administrators to alter the pool of test-takers.[32] Each year, a certain number of students do not take the ITBS either because they are absent on the exam day or because they are exempt from testing due to placement in certain

[31] For eighth grade, we compare achievement in the 1996 and 1998 cohorts in order (i) to compare scores on comparable test forms and (ii) to avoid picking up test score gains due solely to increasing familiarity with a new exam. There is a considerable literature showing that test scores increase sharply the second year an exam is given because teachers and students have become more familiar with the content of the exam. See Koretz (1996). For fourth grade, we do not use the 1998 cohort because of the compositional changes due to third grade retentions in 1997. Instead, we compare achievement gains from 1996 to 1997.

bilingual or special education programs. Other students in bilingual or special education programs are required to take the ITBS but their scores are not reported, meaning that they are not subject to the social promotion policy and their scores do not contribute to the determination of their school's probation status. Under the probation policy, teachers have an incentive to dissuade low-achieving students from taking the exam and/or to place low-achieving students in bilingual or special education programs so that they do not need to take the ITBS.[33] Similarly, teachers may also have an incentive to retain students prior to the promotional gate grades in order to provide additional instruction for the students and thereby reduce retention rates in the more highly publicized gate grades.

Figure 5 shows trends in the proportion of students who were (a) tested with scores reported and (b) in special education. The sample only includes third, sixth and eighth grade students from 1994 to 2000 because some special education and reporting data is not available for the 1993 cohort. Bilingual students are excluded from this analysis since changes in the bilingual policy are confounded with the introduction of high-stakes testing. The top panel shows that the percent of students who were tested and included for reporting purposes has declined steadily since 1994, particularly in the sixth and eighth grades. More importantly, it

---

[32] There is no evidence that the accountability policy has affected the probability of elementary students transferring to private schools, moving out of the district or dropping out of school. Figures available from the author upon request.

[33] Schools are not explicitly judged on the percentage of their students who take the exams, although it is likely that a school with an unusually high fraction of students who miss the exam would come under scrutiny by the central office. In a recent descriptive analysis of testing patterns in Chicago, Easton et al. (2000, 2001) found that the percent of ChiPS students who are tested and included for reporting purposes declined during the 1990s, although they attribute this decline to an increase in bilingual students in Chicago along with changes in the bilingual testing policy. Prior to 1997, the ITBS scores of all bilingual students who took the standardized exams were included for official reporting purposes. During this time, ChiPS testing policy required students enrolled in bilingual programs for more than three years to take the ITBS, but teachers were given the option to test other bilingual students. According to school officials, many teachers were reluctant to test bilingual students, fearing that their low scores would reflect poorly on the school. Beginning in 1997, ChiPS began excluding the ITBS scores of students who had been enrolled in bilingual programs for three or fewer years to encourage teachers to test these students for

appears that the trend became steeper beginning in 1997, suggesting that the accountability policy may have influenced teacher and administrator behavior. Similarly, we see that the proportion of students receiving special education services increased sharply for sixth and eighth graders beginning in 1997 and for third graders in 1999.

Table 11 shows the corresponding Probit estimates for special education placement (the cells show the marginal effects evaluated at the mean). The sample is limited to the 1994-1998 cohorts because estimates for the later cohorts may be confounded by earlier grade retention.[34] Controls include demographics, prior achievement, prior testing status and prior special education placement as well as a pre-existing trend (estimated off of the 1994-1996 cohorts). Column 1 shows the estimates for the full sample. The results suggest that the accountability has increased the proportion of students receiving special education services between 1 and 3 percentage points by 1998, which translates to relative increases of 14 to 24 percent. The next three columns show that these effects are concentrated in the lowest-achieving schools.

The final three columns of Table 11 show the estimates separately by school achievement level, but only for those students whose prior achievement put them at risk for special education placement (i.e., students in the bottom quartile of the national achievement distribution). Notice that the top performing schools were more aggressive in placing students in special education prior to the accountability policy, perhaps because these students were performed lower relative to school average achievement level and were thus more obvious candidates for evaluation. Here

---

diagnostic purposes. In 1999, the ChiPS began excluding the scores of fourth year bilingual students as well, but also began requiring third-year bilingual students to take the ITBS exams.

[34] Students who were previously in special education were more likely to have received waivers from the accountability policy, and thus more likely to appear in the 1999 or 2000 cohorts. One alternative would be to control for special education placement at *t-3* or *t-4*, but data is not available this far back for the earlier cohorts.

we see that the highest risk students, conditional on their prior achievement level,[35] were more likely to be placed in special education under the accountability regime if they were attending low-achieving schools. For example, the lowest performing schools increased special education placements for high-risk sixth graders by 50 percent following the introduction of the accountability policy, compared with an increase of roughly 32 percent among moderate-achieving schools and no increase among the highest performing schools. This is consistent with the incentives provided by the policy.

7.3 Grade retention

Another way for teachers to shield low-achieving students from the accountability mandates is to preemptively retain them—that is, hold them back before they enter grade three, six or eight. By doing so, teachers allow these children to mature and gain an additional year of learning before moving to the next grade and facing the high-stakes exam. Thus, even in grades not directly affected by the promotional policy retention rates may have increased under high-stakes testing.[36] However, because teachers (and parents) are extremely reluctant to retain students multiple times, one would predict retention rates in grades four, five and seven to increase initially, but then level off or decline as the probability that students entering these grades have already been retained once in an earlier grade increases.[37] Figure 6 shows this exact pattern. Prior to the accountability policy, the retention rate was roughly 4 to 5 percent in first

[35] We have controlled for the students prior achievement level in each regression using third order polynomials in prior reading and math.

[36] Roderick et al. (2000) found that retention rates in kindergarten, first and second grades started to rise in 1996 and jumped sharply in 1997 among first and second graders. Building on this earlier work, the analysis here (a) controls for changes in student composition and pre-existing trends, (b) explicitly examines heterogeneity across students and (c) examines similar trends in grades four, five and seven.

[37] Alternatively, one would predict a cumulative measure of grade retention by any point in time to increase more consistently, perhaps level off, but certainly not decline.

grade, 2.5 percent in second grade and a little over 1 percent in grades four, five and seven. Retention rates began to increase in 1996, possibly in anticipation of the new standards the students would face in 1997. In most grades, the rates peaked in 1997 and then declined somewhat. However, the first grade retention rate continued to increase over time. This is consistent with the fact that first grade is likely the first opportunity for retention for most students, while teachers in other grades take prior retentions into consideration when in deciding whether or not to hold a student back.

Table 12 presents Probit estimates of the effect of high-stakes testing on grade retention in these grades. The dependent variable is a binary indicator that takes on the value one if the student was enrolled in the same grade the following year, and zero otherwise. The top panel replicates the trends shown in Figure 6, but also controls for student, school and neighborhood demographics. In comparison to 1993-95, retention rates in 1997 increased by 33 percent in first grade, 100 percent in second grade and 150-200 percent in grades four, five and seven. The bottom panel controls for current achievement, age and special education status as well as demographic variables, thereby accounting for prior retention and giving a better sense of the marginal effect of the policy on the propensity to retain students. Notice that the estimates for 1997 and 1998 do not change much, but the estimates for 1999 and 2000 increase somewhat.

7.4 Sensitivity analysis

To test the sensitivity of the findings presented in the previous sections, Table 13 presents comparable estimates for a variety of different specifications and samples. For simplicity, I only present result for the 1998 eighth grade cohort. (The sensitivity results are comparable for the other grades and cohorts. Tables available from author upon request.) Row 1 shows the baseline

estimates. The next three rows show that the results are not sensitive to including students who either were in that grade for the second time (e.g., retained students) or whose test scores were not included for official reporting purposes because of a special education or bilingual classification. Rows 5 and 6 expand the sample even further, including students with missing outcome data, and instead imputing test scores using different rules. The inclusion of these students does not change the results. Rows 7 to 9 examine the robustness of the findings to the exclusion of prior test score data and/or pre-existing achievement trends, finding that neither of these alternative specifications substantially change the results. Row 10 presents results that include school fixed effects and obtain similar results, indicating that the composition of schools in Chicago did not change appreciably over this time period. Finally, rows 11 and 12 estimate the findings using only the 1994, 1996 and 1998 cohorts, all of which took Form L of the ITBS. This should control for any changes in form difficulty that may confound the results. We see that while the results shrink somewhat, they are still statistically significant and large in magnitude.

8. Conclusions

When the federal legislation *No Child Left Behind* became law earlier this year, high-stakes testing took on a heightened level of importance for students, teachers and parents across the country. The results of this analysis suggest that HST substantially increases math and reading performance, with test score gains on the order of 0.20 to 0.30 standard deviations. To put these results in perspective, it is useful to compare them to other education programs.[38] One of the most popular reform strategies that has been shown to improve student achievement is

[38] This is complicated by the fact that there is little compelling evidence that many popular education reform strategies, such as raising teacher salaries or increasing certification requirements, have *any* impact on student achievement (Hanushek 1996).

reducing class sizes. Results from a randomized experiment, Tennessee STAR, suggests that reducing class size in the early elementary grades from 22 to 15 students raises achievement by roughly 0.20 standard deviations (Krueger 1999, Nye et. al. 1999, Finn and Achilles 1999).

If the benefits of HST are equal to or greater than most other education programs, the costs are almost certainly lower. Based on an analysis of school accountability systems throughout the country, Hoxby (2001) concludes that the current state accountability programs cost between $5 and $35 per pupil each year. These figures include the costs of assessment (e.g., writing and publishing standards, purchasing and scoring exams, publishing results, and designing/piloting new assessments if off-the-shelf exams are not used) as well as the cost of running an office of accountability (e.g., increased staff to promulgate standards, run seminars for teachers and principals, answer questions for parents, calculate and monitor school progress, assist failing schools, etc.). California and Texas, states with relatively sophisticated and comprehensive accountability systems, spend only $20 per pupil per year on their programs, which amounts to less than 0.3 percent of total per pupil expenditures in these states. In comparison, Hoxby (2001) estimates that a 10 percent reduction in class size (about 2 kids per classroom) would cost roughly $615 per pupil and a 10 percent increase in teacher compensation would cost roughly $437 per student. In a cost-benefit analysis of STAR, Krueger (2000) estimates that class size reductions of the magnitude examined in this randomized experiment would cost $3,501 per pupil each year a student is in a small class.[39]

[39] This assumes that the cost of creating and staffing new classrooms is proportional to the annual per pupil cost, and is based on the 47 percent increase in classrooms implied by the class size reductions in STAR (7/15=0.467) and uses the 1997-98 national average per pupil expenditure figure of $7,502. Using the national average per pupil expenditures in 2000-2001 of $8,157, the cost would be $3,807. Hoxby (2001) estimates that the cost of class size reductions is proportional to the proportion of per pupil expenditures devoted to teacher compensation and other costs that are proportional to building size, which she estimates as roughly 74 percent of per pupil expenditures. Based on these assumptions, the cost of class size reductions similar to those in STAR would be $2,817 in 2000-

While the Chicago accountability program was somewhat more extensive than others, it too appears to be relatively inexpensive. The annual assessment and administrative cost of the accountability program in Chicago is roughly $13 per pupil.[40] Unlike most state systems, the Chicago program included a number of support programs for students and schools. In 2000-2001, Chicago spent $43.7 million on the summer school program and $12 million on the Lighthouse afterschool program. These services for low-achieving students amounted to roughly $144 per pupil.[41] The increases in special education placement also imposed a cost. If we make the conservative assumption that special education rates increased by two percentage points in *all* grades (mirroring the increases we saw in grades three, six and eight), this would translate to an additional expenditure of $40 per pupil.[42] Finally, the policy of ending social promotion had a large potential impact on the cost of the program. During the first four years of the accountability policy, roughly 5 percent of all elementary students were retained each year, which translates to roughly 3 percent of all ChiPS students.[43] If we assume that each of these students will remain in school an additional year, the annual cost of ending social promotion would be

---

2001. These estimates do not take into account the potential decline in teacher quality that may result from wide-scale reduction in class size.

[40] Based on expenditures for the Office of Accountability of $5.2 million in 2000-2001 and average daily attendance of 387,000 students.

[41] Because the analysis presented in this paper does not capture many of the benefits of these programs, including the full cost of the programs will tend to overstate the overall cost of the accountability policy. For example, the Lighthouse after-school program is targeted largely at students who have been retained. By focusing on first-time or non-retained students, the analysis above would not capture much of the benefit associated with the program. Similarly, many of the students who attended summer school are not captured in this analysis. For example, third grade students who attended summer school in 1997 (and passed in August) would be in the 2000 cohort of 6th graders, the last to be included in the sample. Thus, the analysis above will not capture the benefits to 3rd grade students who attended summer school in 1998, 1999 or 2000. For a separate analysis of summer school, that examines the full causal impact of the program, see Jacob and Lefgren (2002a).

[42] This assumes that per pupil expenditures are roughly 1.25 times greater for special education students with mild learning disabilities compared with regular education students (Chambers 1998) and is based on the per pupil expenditures in Chicago in 2000-2001 of $8,047.

[43] This combines the 7 to 15 percent retained in grades 3, 6 and 8 with the small increases in preemptive retentions in other elementary grades.

roughly $250 per pupil.[44] Thus, a conservative estimate of the total cost of the accountability policy in Chicago is roughly $447 per pupil, still a fraction of the cost of a class size reduction comparable to STAR.

While test-based accountability appears to improve student achievement at a relatively low cost, it also has several potential drawbacks. Insofar as the test score gains were driven largely by an improvement in certain specific skills and/or student effort, it is likely that they will not generalize well to alternative performance measures, particularly those that tap other domains of knowledge. The accountability policy also led to modest increases in special education placement and grade retention. There is little current evidence on the long-term effects of these practices, though many educators content that they have negative consequences for students.[45]

The passage of the *No Child Left Behind* ensures that test-based accountability will be a pervasive force in elementary and secondary education for years to come. This study provides some of the first credible empirical evidence such policies. I find that the accountability policy in Chicago led to substantial increases in math and reading achievement, driven largely by an increase in certain test-specific skills and student effort. In addition, I find that teachers respond strategically to the policy along a variety of other dimensions, most importantly by placing marginal students in special education programs where their scores are not reported for

[44] This figure likely overstates the cost of ending social promotion for three reasons: (1) given the fact that graduation rates in Chicago are roughly 60%, it is likely that many of the retained students will drop out of school, in which case retention may not increase the total years of schooling per student; (2) these costs should be discounted because the school system will not incur the costs of additional schooling for many years; (3) because early grade retention greatly reduces the probability of later retention, annual steady state retention rate for elementary students will likely be lower than 5 percent.

[45] Unfortunately, there is little good evidence on the long-term causal impact of either intervention. Hanushek et. al. (1998) find that special education has a modest positive impact on achievement. Jacob and Lefgren (2002a) find that grade retention has a mixed effect on achievement. Other studies suggest that grade retention increases the likelihood of dropping out, although this research is plagued by selection bias.

accountability purposes. Overall, these results suggest that high-stakes testing has the potential to substantially improve student learning, but needs to be approached with caution.

References

Ashenfelter, O. (1978). "Estimating the Effect of Training Programs on Earnings." The Review of Economics and Statistics. 60(1): 47-57.

Bishop, J. (1998). Do Curriculum-Based External Exit Exam Systems Enhance Student Achievement? Philadelphia, Consortium for Policy Research in Education, University of Pennsylvania, Graduate School of Education**:** 1-32.

Cannell, J. J. (1987). Nationally Normed Elementary Achievement Testing in America's Public Schools: How All Fifty States Are Above the National Average. Daniels, W. V., Friends for Education.

Center for Education Reform (2002). National Charter School Directory. Washington, D.C.

Chambers, Jay G. (1998). "The Patterns of Expenditures on Students with Disabilities: A Methodological and Empirical Analysis," in T. Parrish, J. Chambers, and C. Guarino, eds., Funding Special Education (Thousand Oaks, CA: Corwin Press, Inc.).

Cullen, Julie Berry and Reback, Randall (2002). "Tinkering Toward Accolades: School Gaming under a Performance Accountability System." Working paper, University of Michigan.

Deere, D. and W. Strayer (2001). "Putting Schools to the Test: School Accountability, Incentives and Behavior." Working paper. Department of Economics, Texas A&M University.

Digest of Educational Statistics (2001). Washington, D.C.: Department of Education.

Easton, J. Q., T. Rosenkranz, et al. (2001). Annual CPS Test Trend Review, 2000. Chicago, IL, Consortium on Chicago School Research.

Easton, J. Q., T. Rosenkranz, et al. (2000). Annual CPS Test Trend Review, 1999. Chicago, Consortium on Chicago School Research.

ECS (2000). ECS State Notes, Education Commission of the States (www.ecs.org).

Figlio, David N. and Getzler, Lawrence (2002). "Accountability, Ability and Disability: Gaming the System?" Working paper. University of Florida.

Figlio, David N. and Winicki, Joshua F. (2002). "Food for Thought? The Effects of School Accountability on School Nutrition." Working paper. University of Florida.

Frederisksen, N. (1994). The Influence of Minimum Competency Tests on Teaching and Learning. Princeton, Educational Testing Services, Policy Information Center.

Grissmer, D. and A. Flanagan (1998). Exploring Rapid Achievement Gains in North Carolina and Texas. Washington, D.C., National Education Goals Panel.

Grissmer, D.W. et. al. (2000). Improving Student Achievement: What NAEP Test Scores Tell Us. MR-924-EDU. Santa Monica: RAND Corporation.

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

Hanushek, E. A. (1996). "School Resources and Student Performance." In Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success, ed. Burtless, G. Washington D.C.: Brookings Institution Press.

Hanushek, Eric A., Kain, John F. and Rivkin, Steven G. (1998). "Does Special Education Raise Academic Achievement for Students with Disabilities." NBER Working Paper #6690.

Hedges, L. V. and R. Greenwald (1996). "Have Times Changed? The Relation between School Resources and Student Performance." In Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success, ed. Burtless, G. Washington D.C.: Brookings Institution Press.

Hoover, H. D. (1984). "The Most Appropriate Scores for Measuring Educational Development in the Elementary Schools: GE's." Educational Measurement: Issues and Practice (Winter): 8-18.

Howell, William G. and Peterson, Paul E. (2002). The Education Gap: Vouchers and Urban Schools. Brookings Institution Press: Washington, D.C.

Hoxby, Caroline M. (2001). "The Cost of Accountability." Working paper. Harvard University.

Jacob, B. A. (2001). "Getting Tough? The Impact of Mandatory High School Graduation Exams on Student Outcomes." Educational Evaluation and Policy Analysis 23(2): 99-122.

Jacob, B. A. and L. Lefgren (2002a). "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." NBER working paper #8918.

Jacob, B. A. and L. Lefgren (2002b). "The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from Reform Efforts in Chicago." NBER working paper #8916.

Jacob, B. A. and S. D. Levitt (2002). "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." Working Paper.

Jacob, B.A. (2002). "Test-Based Accountability and Student Achievement Gains: Theory and Evidence." Working paper.

Klein, S. P., L. S. Hamilton, et al. (2000). What Do Test Scores in Texas Tell Us? Santa Monica, CA, RAND.

Koretz, Daniel (1996). Using Student Assessments for Educational Accountability. In Hanushek, Eric A. and Jorgensen, Dale W. (eds.) Improving America's Schools: The

Role of Incentives. Washington, D.C.: National Academy Press. (Chapter 9, pages 197-223.)

Koretz, D., R. L. Linn, et al. (1991). The Effects of High-Stakes Testing: Preliminary Evidence About Generalization Across Tests. American Educational Research Association, Chicago.

Koretz, D. M. and S. I. Barron (1998). The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS). Santa Monica, RAND.

Krueger, Alan B. (2000). "Economic Considerations and Class Size." Working Paper #447. Princeton University, Industrial Relations Section.

Ladd, H. F. (1999). "The Dallas School Accountability and Incentive Program: An Evaluation of its Impacts on Student Outcomes." Economics of Education Review **18**: 1-16.

Linn, R. L., M. E. Graue, et al. (1990). "Comparing State and District Results to National Norms: The Validity of the Claim that 'Everyone is Above Average'." Educational Measurement: Issues and Practice **9**(3): 5-14.

National Center for Education Statistics (1997). "Public School Choice Programs, 1993-94: Availability and Student Participation." Issue Brief NCES 97-909. Washington, D.C.: Department of Education.

Neill, M. and K. Gayler (1998). Do High Stakes Graduation Tests Improve Learning Outcomes? Using State-Level NAEP Data to Evaluate the Effects of Mandatory Graduation Tests. High Stakes K-12 Testing Conference, Teachers College, Columbia University.

Pearson, D. P. and T. Shanahan (1998). "The Reading Crisis in Illinois: A Ten Year Retrospective of IGAP." Illinois Reading Council Journal **26**(3): 60-67.

Petersen, N. S., Kolen, M. J. and Hoover, H. D. (1989). "Scaling, Norming and Equating." In Handbook of Educational Measurement (3rd edition). 221-262.

Richards, Craig E. and Sheu, Tian Ming (1992). The South Carolina School Incentive Reward Program: A Policy Analysis. Economics of Education Review 11(1): 71-86.

Robelen, Erik W. (2002). An ESEA Primer. *Education Week*. February 21, 2002.

Roderick, M. and M. Engel (2001). "The Grasshopper and the Ant: Motivational Responses of Low-Achieving Students to High-Stakes Testing." Educational Evaluation and Policy Analysis. 23(3): 197-227.

Roderick, M., J. Nagaoka, et al. (2000). Update: Ending Social Promotion. Chicago, IL, Consortium on Chicago School Research.

Shepard, L. A. (1990). "Inflated Test Score Gains: Is the Problem Old Norms or Teaching the Test?" Educational Measurement: Issues and Practice **9**(3): 15-22.

Smith, S. S. and R. A. Mickelson (2000). "All that Glitters is Not Gold: School Reform in Charlotte-Mecklenburg." Educational Evaluation and Policy Analysis 22(2): xxx.

Stecher, B. M. and S. I. Barron (1999). Quadrennial Milepost Accountability Testing in Kentucky. Los Angeles, Center for the Study of Evaluation, University of California.

Tepper, R. L. (2002). The Influence of High-Stakes Testing on Instructional Practice in Chicago. Doctoral dissertation. Harris Graduate School of Public Policy, University of Chicago.

Tepper, R.L., Stone, S. and Roderick, M. (Forthcoming). Ending Social Promotion: The Effects on Teachers and Students. Consortium on Chicago School Research. Chicago, IL.

Toenjes, L. Dworkin, A. G. Lorence, J. and A. N. Hill (2000). "The Lone Star Gamble: High Stakes Testing, Accountability and Student Achievement in Texas and Houston." Mimeo.

The Sociology of Education Research Group (SERG), Department of Sociology, University of Texas.

Winfield, L. F. (1990). "School Competency Testing Reforms and Student Achievement: Exploring a National Perspective." Educational Evaluation and Policy Analysis **12**(2): 157-173.

Table 1: Summary Statistics

Variables	Low-Stakes (1993-1996)	High-Stakes (1997-2000)
Student Outcomes		
Tested[a]	0.958	0.962
Tested and Scores Reported[a]	0.866	0.839
In Special Education	0.116	0.139
ITBS Math Score (GE's relative to national norm)[b]	-0.76	-0.25
ITBS Reading Score (GE's relative to national norm)[b]	-0.96	-0.58
Accountability Policy[c]		
Percent who failed to meet promotional criteria in May	--	0.393
Percent retained or in transition center next year	--	0.078
Percent attending school on academic probation	--	0.108
Student Demographics		
Prior math achievement (GE's relative to national norm)[d]	-0.58	-0.42
Prior reading achievement (GE's relative to national norm)[d]	-0.89	-0.71
Male	0.505	0.507
Black	0.544	0.536
Hispanic	0.305	0.326
Age[b]	11.839	11.719
Living in foster care	0.032	0.051
Free or reduced price lunch	0.795	0.861
In bilingual program (currently or in the past)	0.331	0.359
Select Neighborhood Characteristics[e]		
Median HH Income	22,700	23,276
% Managers/Professionals (of those working)	0.169	0.169
Poverty Rate	0.269	0.254
% not working	0.407	0.402
Female Headed HH	0.406	0.391
Number of observations	370,210	397,057

Notes: The sample includes students in grades 3, 6 and 8 from 1993 to 2000 who were not missing demographic information. [a] Excludes bilingual students. [b] Excludes retainees (i.e., students attending the grade for the second or third time). [c] Includes students in 1997 to 2000 cohorts, although the promotional criteria changed somewhat over this period. [d] Excludes students in grade three since sufficient prior achievement measures were not available. [e] Based on the census tract in which the student was living, with data taken from the 1990 census.

Table 2: OLS Estimates of ITBS Math and Reading Achievement

	Dependent Variable: Standardized ITBS Score	
3rd Grade	***Reading***	***Math***
2000 Cohort	0.186 (0.033)	0.263 (0.037)
1999 Cohort	0.212 (0.028)	0.190 (0.031)
1998 Cohort	0.173 (0.019)	0.213 (0.021)
1997 Cohort	0.026 (0.018)	-0.081 (0.019)
6th Grade		
2000 Cohort	0.161 (0.022)	0.326 (0.027)
1999 Cohort	0.118 (0.018)	0.154 (0.023)
1998 Cohort	0.212 (0.014)	0.243 (0.017)
1997 Cohort	0.085 (0.012)	0.088 (0.014)
8th Grade		
2000 Cohort	0.240 (0.024)	0.459 (0.026)
1999 Cohort	0.192 (0.021)	0.485 (0.022)
1998 Cohort	0.197 (0.015)	0.306 (0.015)
1997 Cohort	0.100 (0.013)	0.318 (0.014)
Includes controls for demographics, prior achievement and pre-existing trends	Yes	Yes

Notes: Includes students in the specified grades from 1993 to 2000. Control variables not shown include race, gender, race*gender interactions, guardian, bilingual status, special education placement, prior math and reading achievement, school demographics (including enrollment, racial composition, percent free lunch, percent with limited English proficiency and mobility rate) and demographic characteristics of the student's home census tract (including median household income, crime rate, percent of residents who own their own homes, percent of female-headed household, mean education level, unemployment rate, percent below poverty, percent managers or professionals and percent who are living in the same house for five years). Prior achievement is measured by math and reading scores three years prior to the base year (i.e., at *t-3*). Missing test scores are imputed using other observable characteristics of the student and a variable is included indicating the score was missing. Second and third-order polynomials in prior achievement are included to account for any non-linear relationship between past and current test scores. Robust standard errors that account for the correlation of errors within schools are shown in parentheses.

Table 3: OLS Estimates of Achievement Trends in Chicago versus Other Large Midwestern Cities

	Dependent Variables			
Independent Variables	Math Score		Reading Score	
Chicago	0.039 (0.056)	-17.94 (63.03)	-0.048 (0.034)	-2.95 (32.95)
1997-2000	-0.022 (0.038)	-0.015 (0.048)	-0.003 (0.023)	-0.032 (0.026)
Chicago*(1997-2000)	0.364 (0.061)	0.330 (0.136)	0.253 (0.037)	0.235 (0.076)
Fixed effects for each district and grade	Yes	Yes	Yes	Yes
Pre-existing trends for Chicago and Other Districts	No	Yes	No	Yes
Number of observations	131	131	131	131

Notes: Observations are district-level averages by grade, subject and year. Scores are standardized using the mean and standard deviation for the earliest available year for that grade and subject. The comparison cities include Cincinnati, Gary, Indianapolis, Milwaukee and St. Louis.

Table 4: Heterogeneity across Student and School Subgroups

	Dependent Variables = ITBS Scores for …					
	Math			Reading		
Independent Variables	3rd Grade	6th Grade	8th Grade	3rd Grade	6th Grade	8th Grade
Model 1						
High-stakes (HS)	0.094 (0.010)	0.153 (0.010)	0.250 (0.013)	0.071 (0.008)	0.156 (0.007)	0.117 (0.010)
Model 2						
High-stakes (HS)	0.070 (0.019)	0.036 (0.018)	0.142 (0.019)	0.008 (0.017)	0.038 (0.015)	-0.015 (0.015)
HS * (Student was < 10th percentile)	-0.006 (0.018)	0.009 (0.016)	-0.110 (0.020)	-0.038 (0.019)	0.001 (0.017)	0.147 (0.020)
HS * (Student was 10-25th percentile)	-0.007 (0.015)	0.027 (0.012)	-0.005 (0.013)	0.032 (0.014)	0.035 (0.013)	0.145 (0.013)
HS* (Student was 26-50th percentile)	-0.002 (0.014)	0.012 (0.010)	0.037 (0.011)	0.055 (0.013)	0.041 (0.011)	0.095 (0.010)
HS * (School had < 20% students scored above the 50th percentile)	0.044 (0.026)	0.159 (0.024)	0.176 (0.034)	0.096 (0.022)	0.144 (0.020)	0.083 (0.026)
HS* (School had 20-40% students scored above the 50th percentile)	0.005 (0.024)	0.081 (0.026)	0.078 (0.027)	0.063 (0.020)	0.079 (0.020)	0.008 (0.020)

Notes: The sample includes first-time, included students in cohorts 1993-1999 for grades three and six, and cohorts 1993-1998 for grade eight. School prior achievement is based on 1995 reading scores. Student prior achievement is based on the average of a student's reading and math score three years earlier for grades six and eight, and one year earlier for grade three. The control variables are the same as those used in Table 2. Robust standard errors that account for the correlation of errors within school are shown in parentheses.

Table 5: Differential Effects of Student versus School Incentives

	Gate Grades (Student accountability in math and reading + school accountability in reading)				**Other Grades** (School accountability in reading)			
	Average	3rd Grade	6th Grade	8th Grade	Average	4th Grade	5th Grade	7th Grade
1997 Cohort								
Math	0.109 (0.009)	-0.081 (0.019)	0.089 (0.014)	0.320 (0.014)	0.125 (0.008)	0.156 (0.015)	0.105 (0.014)	0.114 (0.013)
Reading	0.070 (0.008)	0.026 (0.018)	0.084 (0.012)	0.100 (0.013)	0.115 (0.008)	0.073 (0.014)	0.137 (0.014)	0.135 (0.012)
1998 Cohort								
Math	0.144 (0.009)	0.155 (0.019)	0.139 (0.014)	0.137 (0.013)	0.185 (0.008)	0.284 (0.014)	0.067 (0.012	0.203 (0.013)
Reading	0.126 (0.008)	0.109 (0.016)	0.156 (0.012)	0.113 (0.014)	0.123 (0.007)	0.142 (0.0142)	0.034 (0.013)	0.194 (0.011)

Notes: The sample includes first-time students who were tested and whose scores were included in reporting. The 1997 estimates come from a model in which prior achievement is measured at *t-3*, as in the baseline results. The 1998 estimates come from a model in which prior achievement is measured at *t-1, t-2 and t-3* in an attempt to account for compositional changes resulting from grade retention in 1997. Robust standard errors that account for the correlation of errors within schools are shown in parenthesis.

Table 6: The Impact of Test-Based Accountability on a Low-Stakes Achievement Exam

	Dependent Variables = Standardized IGAP Scores for …					
	Math			Reading		
Specification	**3rd Grade**	**6th Grade**	**8th Grade**	**3rd Grade**	**6th Grade**	**8th Grade**
Sample: Chicago + Comparison Districts (independent variable is interaction between Chicago and high-stakes testing)						
(1) No controls	0.277 (0.030)	0.280 (0.030)	0.229 (0.047)	0.313 (0.030)	0.246 (0.030)	0.269 (0.043)
(2) Controlling for time-varying school and district characteristics	0.123 (0.066)	0.171 (0.048)	0.054 (0.052)	0.176 (0.061)	0.134 (0.046)	0.149 (0.057)
(3) Controls + District specific trends from 1993 to 1996	-0.146 (0.050)	-0.113 (0.040)	-0.061 (0.066)	-0.113 (0.041)	-0.180 (0.055)	0.016 (0.086)
Sample: Chicago alone (independent variable is indicator for high-stakes testing)						
(4) No controls	0.403 (0.033)	0.435 (0.036)	0.397 (0.039)	0.224 (0.030)	-0.159 (0.035)	-0.204 (0.042)
(5) Controls + District specific trends from 1993 to 1996	-0.217 (0.042)	-0.120 (0.040)	0.028 (0.050)	-0.246 (0.035)	-0.191 (0.038)	0.187 (0.048)

Notes: The following control variables are also included in the regressions shown above: percent black, percent Hispanic, percent Asian, percent Native American, percent low-income, percent Limited English Proficient, average daily attendance, mobility rate, school enrollment, pupil-teacher ratio, log(average teacher salary), log(per pupil expenditures), percent of teachers with a BA degree, and the percent of teachers with a MA degree or higher. Robust standard errors that account for correlation within schools across years are shown in parenthesis. The regressions are weighted by the inverse square root of the number of students enrolled in the school.

Table 7: The Impact of Test-Based Accountability on Low-Stakes Achievement Test Scores, by School Prior Achievement

	Dependent Variables = Standardized IGAP Scores for …					
	Math			Reading		
Specification	**Bottom Schools**	**Middle Schools**	**Top Schools**	**Bottom Schools**	**Middle Schools**	**Top Schools**
Sample: Chicago alone (independent variable is indicator for HST years, including controls + trends)	-0.020 (0.031)	-0.045 (0.034)	-0.142 (0.063)	-0.047 (0.031)	0.006 (0.031)	-0.125 (0.057)
Sample: Chicago + Comparison Districts (independent variable is interaction between Chicago and HST years, including controls + trends)	0.050 (0.115)	-0.053 (0.050)	-0.105 (0.070)	0.318 (0.121)	-0.025 (0.057)	-0.058 (0.051)

Notes: In the first specification, schools are categorized on the basis of their 1995 ITBS reading scores as described in Table 4. Bottom schools had fewer than 20 percent of students meeting national norms in reading, middle schools had between 20 and 40 percent of students meeting national norms, and top schools had more than 40 percent at this level. In the second specification, schools are categorized into three equal size groups on the basis of their IGAP scores in the early 1990s (because few districts outside Chicago take the ITBS and district specific achievement data is not provided on the ISBE report cards). The regressions include all of the control variables described in Table 7. The regressions are weighted by the inverse square root of the number of students enrolled in the school.

Table 8: The Relationship between Item Type, Position and Improvement on the ITBS Math Exam

	Dependent Variable = Proportion of Students Answering the Item Correctly on the ITBS Math Exam		
Independent Variables	**(1)**	**(2)**	**(3)**
1998 Cohort	0.017 (0.011)	0.015 (0.014)	0.035 (0.013)
Basic Skills * 1998	0.022 (0.005)		
Math Computation *1998		0.025 (0.008)	
Whole numbers			
Decimals			0.000 (0.010)
Fractions			0.022 (0.017)
Number Concepts *1998		0.023 (0.008)	
Equations and inequalities			0.002 (0.015)
Fractions, decimals, percents			0.004 (0.013)
Geometry			0.002 (0.013)
Measurement			-0.028 (0.016)
Numeration and operations			0.001 (0.011)
Probability and statistics			0.011 (0.018)
Other Skills * 1998			
Estimation *1998		0.003 (0.012)	
Compensation			-0.043 (0.012)
Order of magnitude			-0.013 (0.015)
Standard rounding			-0.002 (0.011)

Data Analysis *1998		0.006 (0.013)	
Compare quantiles			-0.018 (0.015)
Interpret relationships			-0.024 (0.012)
Read amounts			-0.002 (0.016)
Problem Solving * 1998			
Multiple step			-0.023 (0.012)
Use strategies			-0.032 (0.014)
Single step			-0.017 (0.013)
2nd Quintile of the Exam * 1998	0.001 (0.001)	0.001 (0.008)	0.002 (0.008)
3rd Quintile of the Exam * 1998	-0.001 (0.008)	-0.002 (0.008)	-0.002 (0.008)
4th Quintile of the Exam * 1998	0.011 (0.008)	0.008 (0.011)	0.008 (0.011)
5th Quintile of the Exam * 1998	0.006 (0.009)	0.003 (0.012)	0.002 (0.012)
Number of Observations	1,038	1,038	1,038
R-Squared	.960	.962	.962

Notes: The sample consists of all tested and included students in grades three, six and eight in years 1994, 1996 and 1998. The units of observation are item*year proportions, reflecting the proportion of students answering the item correctly in that year. Fixed effects for grade, main effects for item difficulty, item difficulty x 1998 and item position are included in the models but not shown here.

Table 9: The Relationship between Item Type, Position and Improvement on the ITBS Reading Exam

	Dependent Variable = Proportion of Students Answering the Item Correctly on the ITBS Reading Exam		
	(1)	(2)	(3)
1998	0.036 (0.028)	0.036 (0.028)	0.045 (0.031)
Construct Factual Meaning * 1998		0.000 (0.009)	
Literal meaning of words			-0.009 (0.020)
Understand factual information			-0.004 (0.014)
Construct Inferential Meaning * 1998		-0.001 (0.009)	
Draw conclusions			-0.009 (0.014)
Infer feelings, traits, motives of characters			0.001 (0.016)
Represent/apply information			-0.003 (0.019)
Construct Evaluative Meaning * 1998			
Author's attitude, purpose, viewpoint			-0.001 (0.018)
Determine main idea			
Interpret non-literal language			-0.008 (0.020)
Structure, mood, style, tone			-0.014 (0.019)
2nd Quintile of the Exam * 1998	0.000 (0.011)	0.000 (0.011)	0.00 (0.011)
3rd Quintile of the Exam * 1998	0.013 (0.012)	0.013 (0.012)	0.013 (0.012)
4th Quintile of the Exam * 1998	0.015 (0.012)	0.015 (0.012)	0.013 (0.012)
5th Quintile of the Exam * 1998	0.031 (0.014)	0.031 (0.014)	0.029 (0.014)
Number of Observations	387	387	387
R-Squared	0.958	0.959	.963

Notes: The sample consists of all tested and included students in grades three, six and eight in years 1994, 1996 and 1998. The units of observation are item*year proportions, reflecting the proportion of students answering the item

correctly in that year. Fixed effects for grade, main effects for item difficulty and item type are included in the models but not shown here. Fixed effects for grade, main effects for item difficulty, item difficulty x 1998 and item position are included in the models but not shown here.

Table 10: Differential Effects on Low versus High Stakes Subjects

	Dependent Variables: ITBS score in …			
Independent Variables	Math	Reading	Science	Social Studies
Model 1				
High-stakes (HS)	0.234 (0.009)	0.172 (0.008)	0.075 (0.008)	0.050 (0.007)
Model 2				
High-stakes (HS)	0.206 (0.017)	0.084 (0.017)	0.074 (0.018)	0.044 (0.018)
HS * (Student was < 10th percentile)	-0.030 (0.023)	0.014 (0.022)	-0.081 (0.022)	-0.069 (0.022)
HS * (Student was 10-25th percentile)	-0.040 (0.017)	0.018 (0.015)	-0.065 (0.017)	-0.058 (0.017)
HS* (Student was 26-50th percentile)	-0.028 (0.014)	0.014 (0.013)	-0.032 (0.015)	-0.029 (0.015)
HS * (School had < 20% students scored above the 50th percentile)	0.083 (0.022)	0.097 (0.020)	0.035 (0.022)	0.030 (0.023)
HS* (School had 20-40% students scored above the 50th percentile)	-0.002 (0.022)	0.056 (0.020)	0.015 (0.022)	0.025 (0.021)

Notes: Cells contain OLS estimates based on comparisons of the 1996 and 1998 cohorts for grade eight and the 1996 and 1997 cohorts for grade four, controlling for the student, school and neighborhood demographics described in the notes to Table 2. ITBS scores are standardized separately by grade and subject, using the 1996 student-level mean and standard deviation. Estimates in the top row are based a model with no interactions. The estimates in the subsequent rows are based on a single regression model that includes interactions between high-stakes testing and student or school prior achievement, with high ability students in high-achieving schools as the omitted category. Robust standard errors that account for the correlations of errors within schools are shown in parentheses.

Table 11: Has High-Stakes Testing Affected Special Education Placement?

		All Students				Students in the bottom quartile of the national achievement distribution		
Grade	**Independent Variables**	All Schools	Bottom Schools	Middle Schools	Top Schools	Bottom Schools	Middle Schools	Top Schools
3rd	Baseline Mean	0.116	0.101	0.122	0.151	0.189	0.279	0.439
	1997 Cohort	0.006 (0.004)	0.016 (0.006)	0.000 (0.007)	-0.019 (0.010)	0.066 (0.018)	0.011 (0.027)	-0.064 (0.073)
	1998 Cohort	0.016 (0.006)	0.026 (0.008)	0.015 (0.011)	-0.018 (0.013)	0.115 (0.024)	0.050 (0.039)	-0.003 (0.096)
6th	Baseline Mean	0.143	0.138	0.146	0.151	0.209	0.276	0.503
	1997 Cohort	0.018 (0.005)	0.021 (0.006)	0.028 (0.009)	-0.003 (0.012)	0.047 (0.014)	0.069 (0.026)	0.039 (0.052)
	1998 Cohort	0.035 (0.007)	0.046 (0.010)	0.042 (0.012)	0.005 (0.016)	0.103 (0.020)	0.088 (0.033)	0.017 (0.066)
8th	Baseline Mean	0.139	0.138	0.145	0.136	0.208	0.287	0.515
	1997 Cohort	0.006 (0.004)	0.016 (0.007)	-0.005 (0.007)	0.000 (0.009)	0.043 (0.016)	-0.012 (0.028)	0.066 (0.052)
	1998 Cohort	0.021 (0.007)	0.031 (0.011)	0.005 (0.010)	0.034 (0.018)	0.075 (0.024)	0.043 (0.036)	0.209 (0.059)
Number of Obs. – 3rd Grade		106,484	55,380	35,388	15,348	19,804	9,095	1,563
Number of Obs. – 6th Grade		94,863	48,737	30,897	14,790	29,484	13,684	2,653
Number of Obs. – 8th Grade		92,766	48,063	29,842	14,129	28,433	12,644	2,366

Notes: All of the estimates above come from Probit models and the marginal effects are shown in the cells. The sample includes all first-time students in these grades from 1994 to 2000. Control variables are the same as those described in the notes to Table 2. Robust standard errors that account for the correlation of errors within schools are shown in parentheses.

Table 12: Has High-Stakes Testing Increased Grade Retention in Grades not Directly Affected by the Social Promotion Policy?

	Dependent Variables = Retained in the same grade the following year				
Sample & Specification	1st Grade	2nd Grade	4th Grade	5th Grade	7th Grade
Controlling for student, school and neighborhood demographics					
1997	0.015 (0.003)	0.024 (0.003)	0.021 (0.003)	0.019 (0.002)	0.025 (0.004)
1998	0.021 (0.004)	0.021 (0.003)	0.016 (0.002)	0.017 (0.002)	0.016 (0.003)
1999	0.027 (0.004)	0.019 (0.003)	0.013 (0.002)	0.007 (0.002)	0.014 (0.003)
2000	0.019 (0.004)	0.015 (0.003)	0.011 (0.002)	0.006 (0.002)	0.010 (0.002)
Controlling for current achievement, age and special education status as well as the demographics from above					
1997	0.017 (0.003)	0.024 (0.003)	0.023 (0.001)	0.020 (0.002)	0.026 (0.003)
1998	0.024 (0.003)	0.022 (0.003)	0.021 (0.002)	0.018 (0.002)	0.018 (0.003)
1999	0.030 (0.004)	0.019 (0.003)	0.018 (0.002)	0.010 (0.002)	0.016 (0.003)
2000	0.023 (0.004)	0.016 (0.003)	0.016 (0.002)	0.009 (0.002)	0.012 (0.003)
Baseline rate (average for 1993-95)	0.046	0.025	0.014	0.012	0.012
Number of observations	273,387	259,240	234,488	227,095	211,905

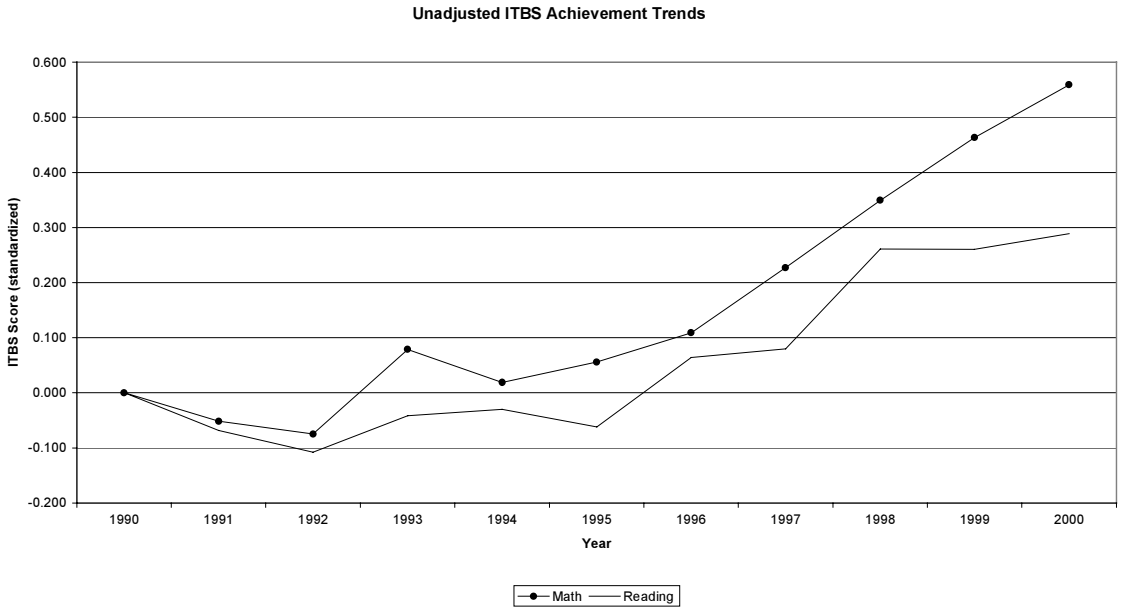
Notes: All of the estimates above come from Probit models and the marginal effects are shown in the cells. Robust standard errors that account for the correlation of errors within school are presented in parentheses. Demographics include gender, race, free lunch, bilingual status, and neighborhood and school characteristics. Current achievement is specified as a second order polynomial in reading and math and current age is specified as a series of dummy variables. The models for first and second graders do not contain any achievement measures since standardized tests are not mandatory until third grade.

Table 13: Sensitivity Analysis

	Dependent Variable	
Specification	ITBS Math Score	ITBS Reading Score
Baseline	0.306 (0.016)	0.197 (0.015)
Including students who were tested, but whose scores were not counted for official reporting purposes (non-reported students)	0.304 (0.016)	0.193 (0.015)
Including students who were in the grade for the second or third time (retained students)	0.309 (0.016)	0.194 (0.015)
Including both non-reported and retained students	0.310 (0.016)	0.193 (0.015)
Including both non-reported and retained students, and imputing scores for students who did not take the ITBS (impute to the 25th percentile of cohort and school)	0.311 (0.013)	0.197 (0.015)
Including both non-reported and retained students, and imputing scores for students who did not take the ITBS (impute to the 10th percentile of cohort and school)	0.321 (0.016)	0.203 (0.015)
No pre-existing achievement trend	0.257 (0.011)	0.177 (0.010)
No controls for prior achievement	0.253 (0.020)	0.138 (0.019)
No controls for prior achievement or pre-existing achievement trends	0.365 (0.012)	0.301 (0.012)
Add school fixed effects	0.299 (0.016)	0.193 (0.015)
Common Form I – only include the 1994, 1996 and 1998 cohorts that all took ITBS Form L (no trend)	0.252 (0.011)	0.164 (0.009)
Common Form II – only include the 1994, 1996 and 1998 cohorts that all took ITBS Form L (with trend)	0.215 (0.019)	0.139 (0.016)

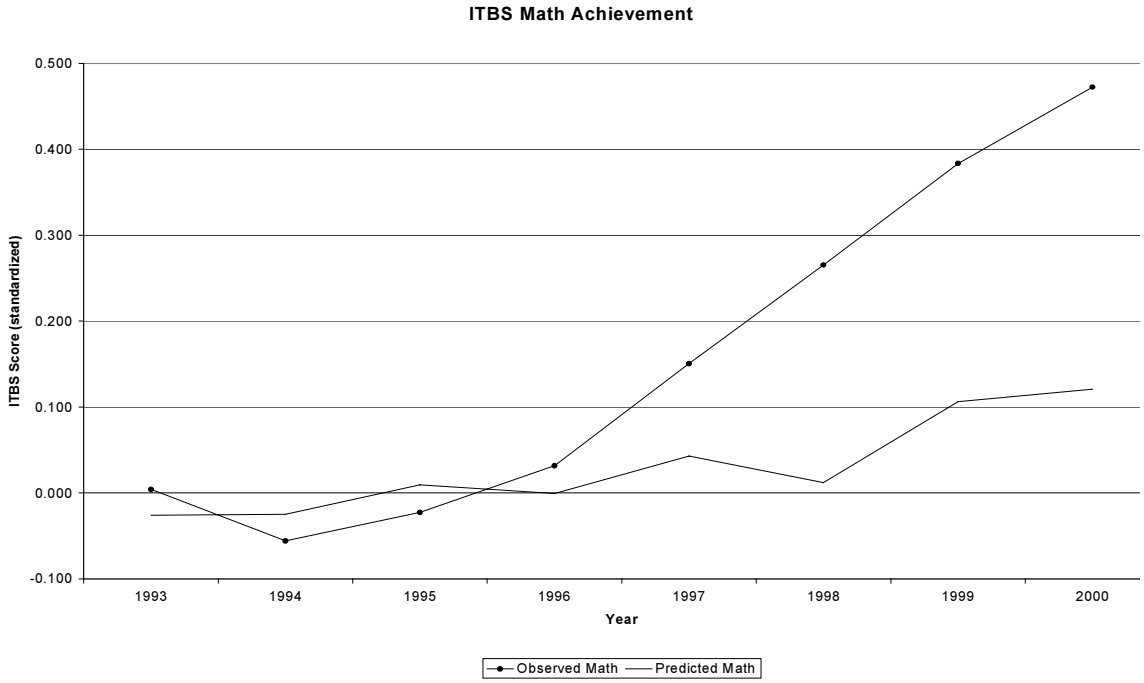
Notes: For the sake of brevity, the estimates shown in the cells above are the effects of high-stakes testing on the 1998 eighth grade cohort. Results are comparable for other grades and cohorts, and are available upon request from the author.

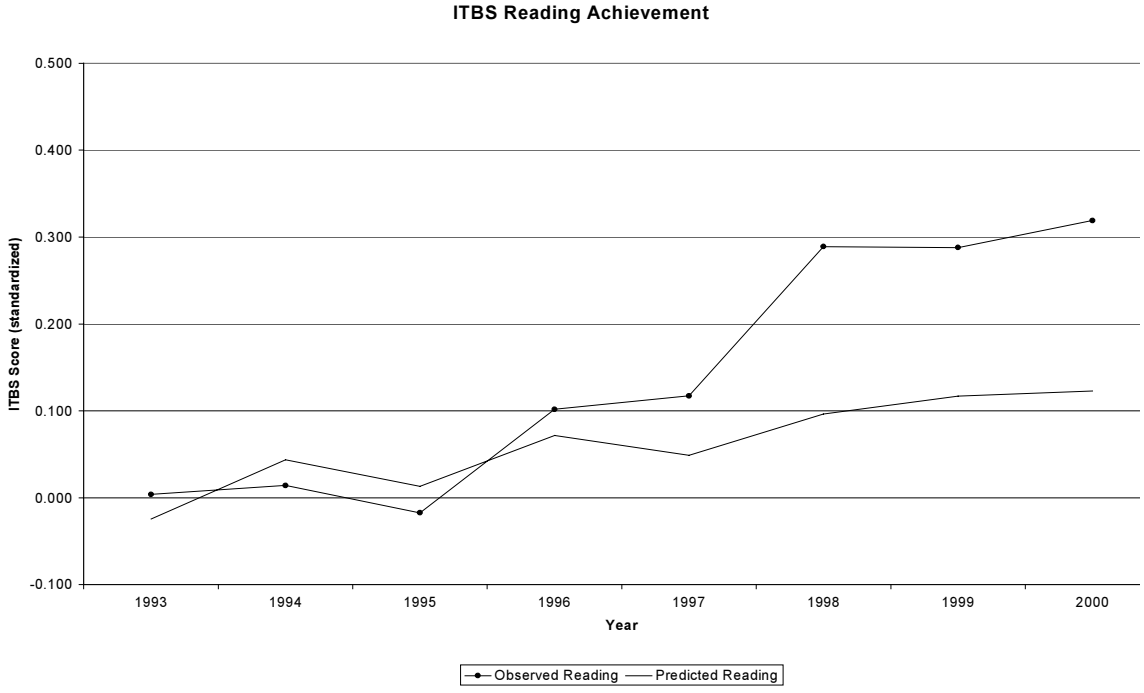
Figure 1: Unadjusted ITBS Achievement Trends in Chicago, 1990-2000



Notes: The sample includes 3rd, 6th and 8th grade students from 1990 to 2000, excluding retainees and students whose scores were not reported. The scores are standardized separately for each grade using the 1990 student-level mean and standard deviation.

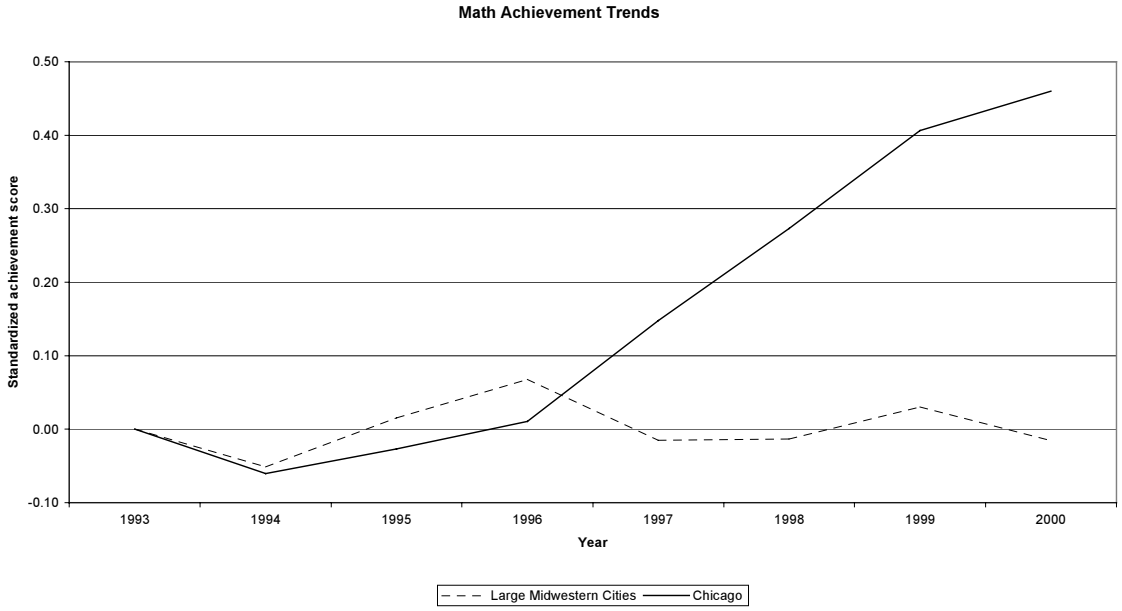
Figure 2: Observed versus Predicted Achievement Levels in Chicago, 1993-2000

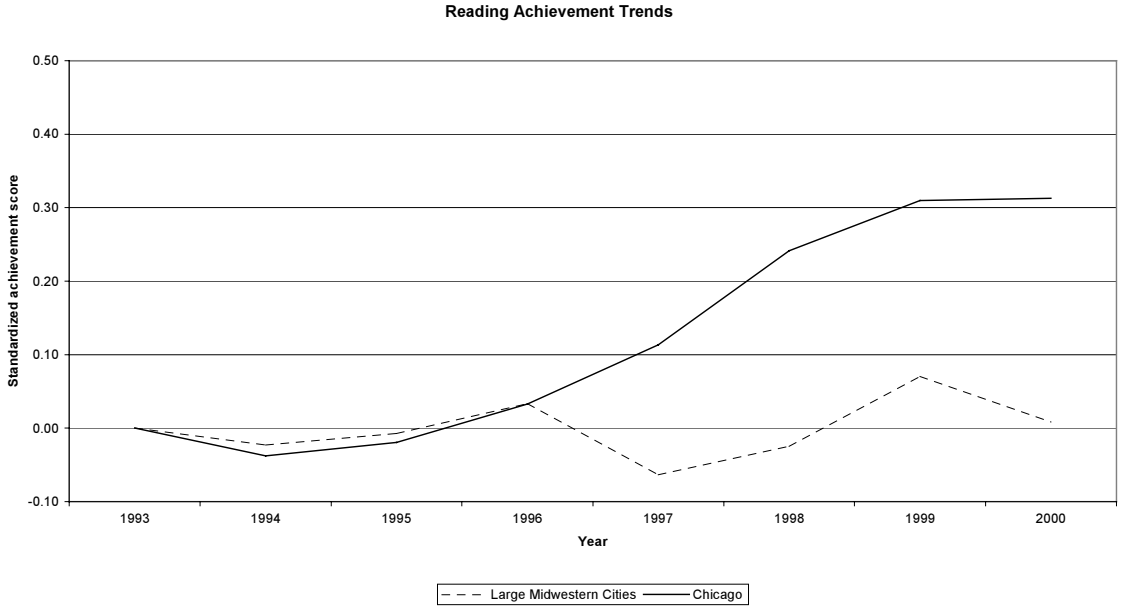




Notes: The sample includes 3rd, 6th and 8th grade students from 1993 to 2000, excluding retainees and students whose scores were not reported. Scores are standardized separately for each grade using the 1993 student-level mean and standard deviation. The predicted scores are derived from an OLS regression on pre-policy cohorts (1993 to 1996) that includes controls for student, school and neighborhood demographics as well as prior student achievement and a linear time trend.

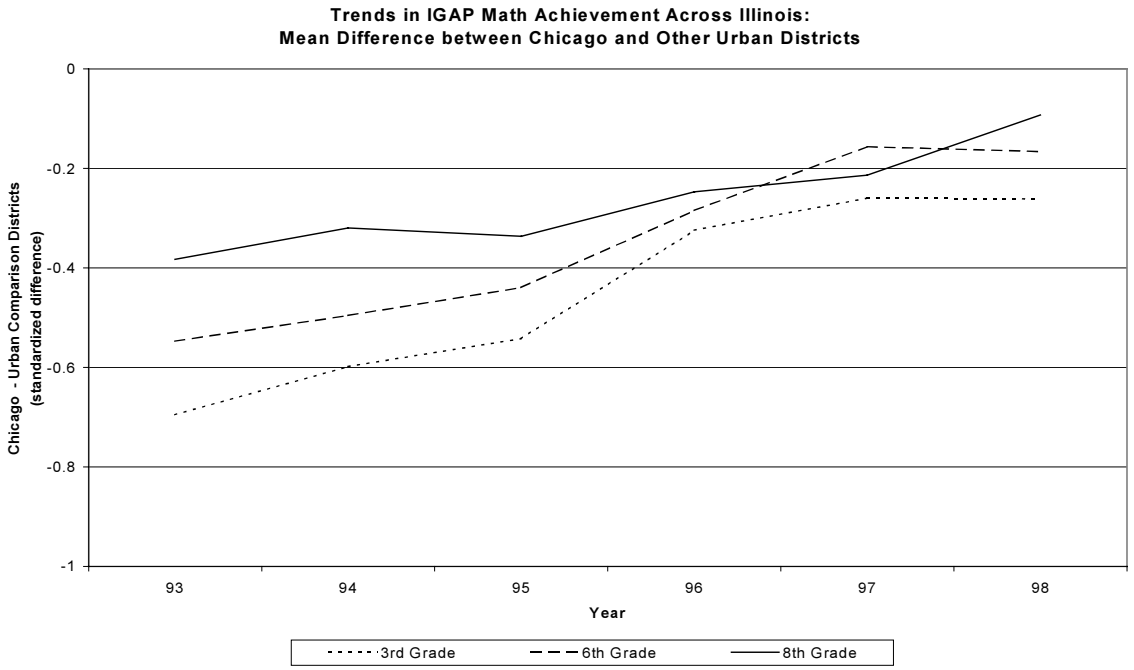
Figure 3: Achievement Trends in Chicago versus Other Large, Urban School Districts in the Midwest, 1990-2000

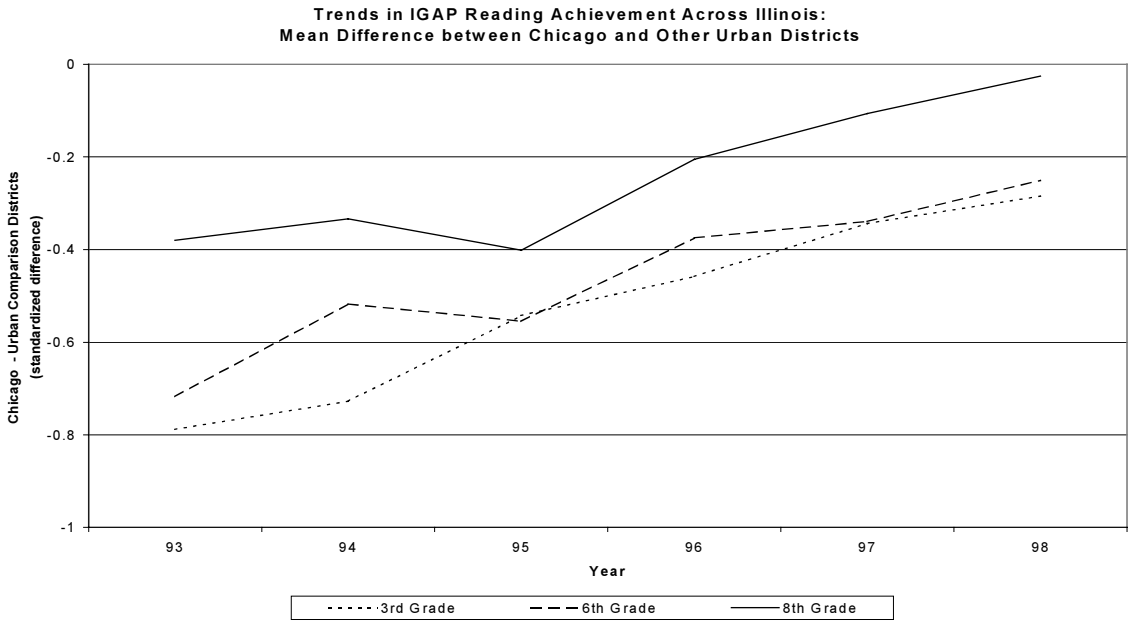




Notes: The achievement series for large Midwestern cities includes data for all tested elementary grades in Cincinnati, Gary, Indianapolis, St. Louis and Milwaukee. The sample includes all grades from 3 to 8 for which test score data was available, and only includes students whose tests scores were reported. Test scores are standardized separately by grade*subject*district, using the student-level mean and standard deviation for the earliest available year.

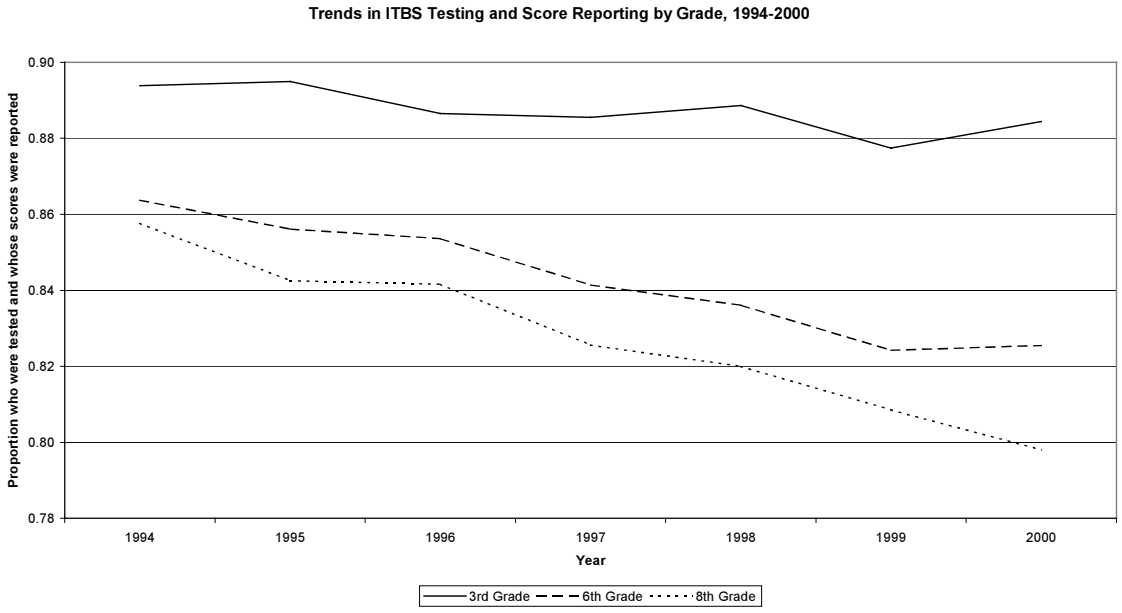
Figure 4: Achievement Trends on Low-Stakes Exam

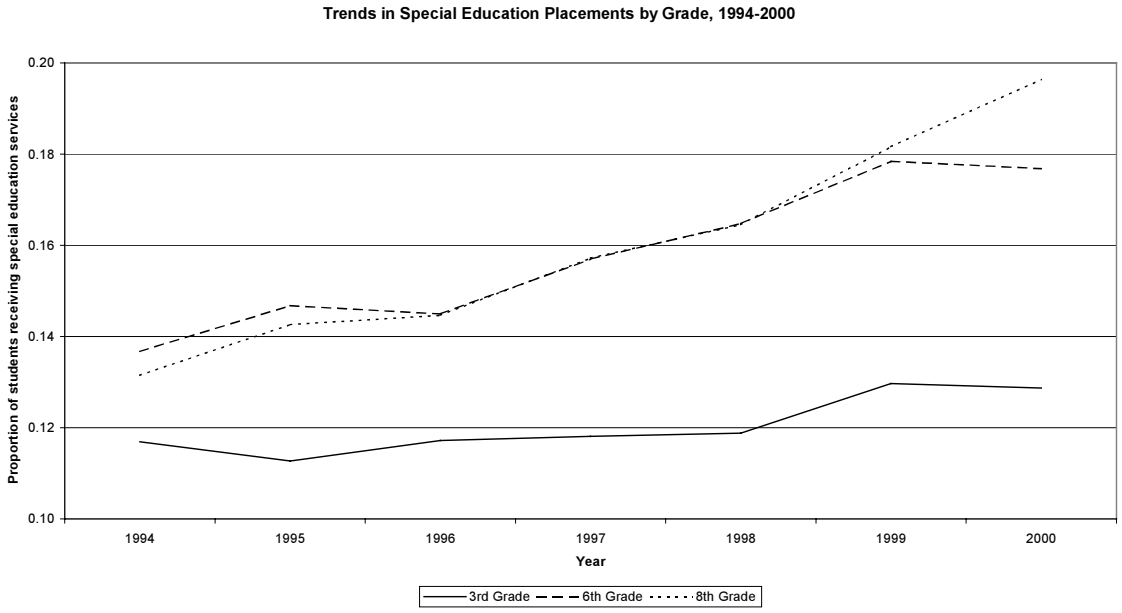




Notes: Chicago averages exclude retained students. District averages are standardized separately using the 1993 state mean and across school standard deviation in the state. The value shown above is the difference in the standardized score for each year. A complete list of the comparison districts can be found in the text.

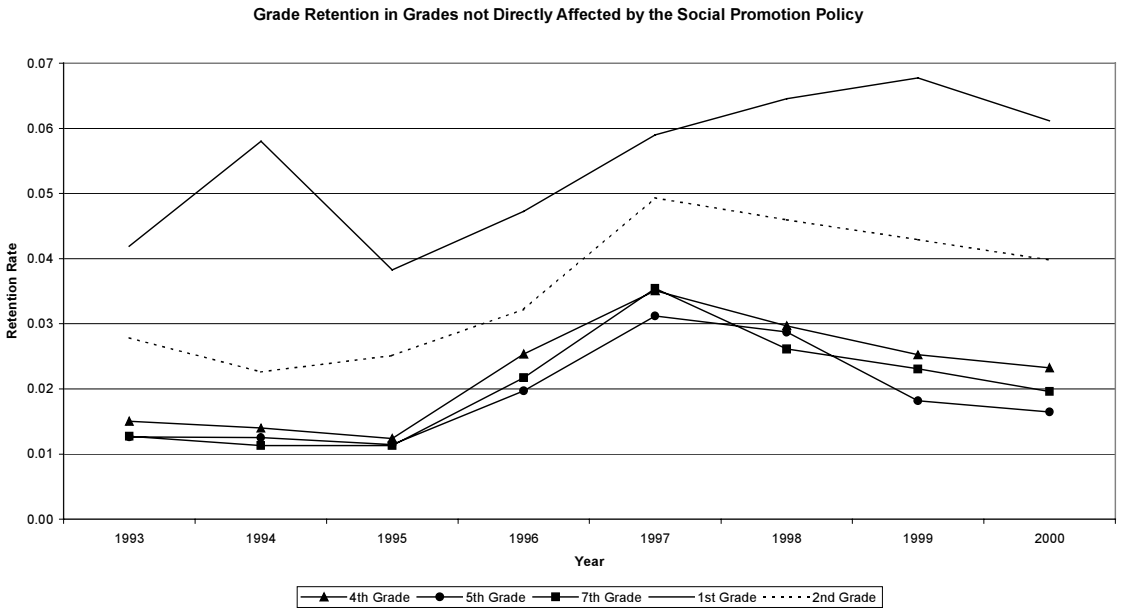
Figure 5: Trends in Testing and Special Education Placements





Notes: The sample includes only first-time, non-bilingual students.

Figure 6: Trends in Grade Retention



Notes: The sample includes only first-time, non-bilingual students.