Preliminary and Incomplete: Comments Welcome

MAKING THE GRADE:
THE IMPACT OF TEST-BASED ACCOUNTABILITY IN SCHOOLS*

Brian A. Jacob
John F. Kennedy School of Government
Harvard University

Original draft: June 2001
This draft: April 2002

Abstract

The recent federal education bill 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, many worry that such incentives will lead to unintended and/or undesirable consequences. This study utilizes detailed administrative data on the Chicago Public School system to examine the impact of a test-based accountability policy on student and teacher behavior. I find that math and reading scores increased sharply following the introduction of a high-stakes accountability policy in Chicago, in comparison to both prior achievement trends in the district and to changes experienced by other large, urban districts in the mid-west. However, I also find evidence that teachers and administrators responded strategically to the incentives along a variety of dimensions. Specifically, the accountability policy led to a substantial increase in the proportion of students placed in special education and to an increase in the proportion of students retained (even in grades not directly affected by the policy). The policy also appears to have led schools to substitute away from low-stakes subjects such as science and social studies. Finally, I show that the accountability policy did not lead to comparable achievement gains on a state-administered, low-stakes exam, suggesting that the gains on the high-stakes exam may have been driven largely by student effort and/or test-specific preparation and thus may not reflect a more general increase in student knowledge.

* 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.

1. Introduction

In January 2002, President Bush signed the "No Child Left Behind" Act of 2001, 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).

School reforms designed to hold students and teachers accountable for student achievement have become increasingly popular in recent years. Statutes in 19 states explicitly link student promotion to performance on a state or district assessment (ECS 2000). The largest school districts in the country, including New York City, Los Angeles, Chicago and Washington, D.C., have recently implemented policies requiring students to attend summer school and/or repeat a grade if they do not demonstrate sufficient mastery of basic skills. At the same time, 20 states reward teachers and administrators on the basis of exemplary student performance and 32 states sanction school staff on the basis of poor student performance. Many states and districts have passed legislation allowing the takeover or closure of schools that do not show improvement (ECS 2000).

While the primary intent of such accountability policies is to provide incentives to maximize student learning, we know that 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 to 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 are focused on secondary students and have little direct impact on teachers or administrators. Recent evidence on school-based accountability programs is mixed. Moreover, these studies generally do not utilize individual student data and thus cannot examine some outcomes of interest or investigate how effects vary across students.

This paper utilizes detailed administrative data to examine the impact of a test-based accountability policy in the Chicago Public Schools (ChiPS).[1] 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, in contrast to earlier studies that 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 incorporated 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.

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

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. This will allow us to begin to separate the effects of student-focused versus school-focused accountability policies.

I find that the accountability policy led to a significant increase in achievement scores, but also induced a variety of other strategic responses on the part of teachers and administrators that might have undesirable consequences for students. Most noticeably, math and reading scores 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. The effects are robust to a variety of different sample and specification choices. This suggests that teachers and students responded quite strongly to the accountability policy. I also find that students in low-achieving schools experienced larger gains than their peers in other schools, consistent with the focus of the school probation policy on low-achieving schools.

I also present evidence that teachers and administrators responded strategically to the incentives along a variety of other dimensions. Following the introduction of high-stakes testing, (i) math and reading scores increased relative to scores on low-stakes exams such as science and social studies; (ii) there was a substantial increase in the proportion of students in special

education and/or excluded from testing; and (iii) retention rates increased substantially in grades not directly affected by the student promotion policy.[2]

Finally, it appears that the accountability policy led to large increases in the high-stakes ITBS scores, but had zero or perhaps even negative effect on a state-administered, low-stakes exam, the Illinois Goals Assessment Program (IGAP). For both exams, students in the lowest performing schools made larger gains relative to their peers in higher performing schools. However, even in low-performing schools, IGAP scores did not deviate significantly from pre-existing trends. There are several potential explanations for the differential achievement trends on the two exams—including differential student effort, ITBS-specific test preparation and cheating. Because factors such as effort, test preparation and cheating are not easily observable, it is difficult to disentangle the reasons underlying the differential achievement patterns. However, I present some evidence suggesting that the extremely large ITBS gains were due at least in part to increased student effort and ITBS-specific preparation. This is consistent with the view that students and teachers substituted toward the high-stakes exam, and raises concerns that the ITBS gains may not reflect a more general increase in student knowledge.

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 and 6 present the main findings and Section 7 concludes.

[2] It is not clear whether these changes resulted from pure gaming on the part of teachers and administrators, or whether they served legitimate education goals as well.

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). The primary drawback of these studies is that 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 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 and 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 also some evidence on strategic responses to test-based accountability. Haney (2000) reports anecdotal evidence that special education placements in Texas have increased under the TAAS program. 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.[3] 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, ending a common practice known as "social promotion" whereby students are advanced to the next grade regardless of their ability or achievement. 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.[4] Students who do not make 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.

One of the most striking features of Chicago's social promotion policy was its scope. 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 affected by the accountability policies. Of those who were subject to the policy, nearly 50 percent of third

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

[4]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

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 eventually 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 are placed on probation. If they do not exhibit sufficient improvement, these schools may be reconstituted, which involves 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.[5] While ChiPS has not reconstituted any elementary schools, teachers and administrators in probation schools report being extremely worried about job security and staff in others report a strong desire to avoid probation.

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:

---

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.

[5] Probation schools received some additional resources and were more closely monitored by ChiPS staff. Jacob and Lefgren (2001b) 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.

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

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.

We face 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).

They found that the additional resources and monitoring provided by probation had no impact on math or reading achievement.

The rich set of longitudinal, student-level data allows one 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.[6] The size and 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.[7]

This strategy has two major drawbacks. First, it does not account for time-varying effects that would have influenced student achievement in a sharp or discontinuous manner. Second, 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

[6] 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.

[7] 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.

example, this may be the case 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:

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

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).

4. Data

This study utilizes detailed administrative data from the ChiPS as well as the Illinois State Board of Education (ISBE). 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.[8] ChiPS personnel and budget files provide information on the financial resources and teacher characteristics in each school and school files provide aggregate

[8] There is no significant change in the percent of 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.

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 – i.e., 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 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.[9] (In section 5.6, I show that the results are robust to changes in the sample and specification.) In order to have sufficient prior achievement data for all students, I limit the analysis to cohorts beginning in 1993.

I delete less than 5 percent of students because they were missing demographic information. In addition, roughly 10 percent of students were not tested each year (most often because of a special education or bilingual placement) and are therefore not included in the

[9] 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.

achievement estimates, although they are included in the estimates of other outcomes. 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 our 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 students were in special education programs and 13 percent of students 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 there were some changes in the student composition during the 1990s. There were slight increases in the percentage of Hispanic students in the ChiPS as well as 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, we see 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, we see 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 have also increased, from 0.116 to 0.139.

5. The Impact of High-Stakes Testing on Student Achievement

5.1 Math and Reading Trends

We begin by examining achievement trends in Chicago over the past eight years. Figure 1 shows the trends in ITBS math and reading scores for grades three, six and eight from 1993 to 2000. Test scores are standardized separately by grade using the 1993 mean and student standard deviation. The predicted values are derived from an OLS regression model that includes cohorts 1993 to 1996 and controls for student, school and neighborhood demographics along with prior academic achievement and a linear time trend. 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 0.45 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.

While there appears to be a sharp increase in achievement following the introduction of HST, Figure 1 also raises some questions. First, the year-to-year achievement scores fluctuate widely at certain points. Second, the timing of the achievement gains is not perfectly correlated with the introduction of the policy. In mathematics it appears that some improvement began as early as 1996. The opposite appears true in reading, where achievement did not begin to substantially increase until 1998.

There are two primary explanations. The first involves changes in the form of the exam across years. The ChiPS administers different forms of the ITBS each year. To the extent that

the forms are not perfectly equated (i.e., one form is more or less difficult than another in a particular grade and/or subject), annual test score changes may not accurately reflect learning gains. This is the reason for some of the choppiness in the trends. Moreover, a new form of the exam was given in 1997, which teachers believe was more difficult than earlier exams, particularly in reading. This may explain why observed reading scores do not increase substantially in 1997. To obtain a cleaner picture of changes in student performance, we can compare cohorts taking similar forms—the ChiPS administered the same form of the exam in 1994, 1996 and 1998 (Form L) and in 1993, 1995 and 2000 (Form K). As I show in Section 5.6, the same picture emerges if we focus on changes across cohorts within form.

Another reason for these trends involves anticipation or implementation effects. To the extent that teachers and students anticipated the change in policy, one might expect improvement to begin prior to 1997. Conversely, if schools found it difficult to quickly implement changes in response to the policy (e.g., changing schedules to shift resources across grades or subjects, hiring new teachers, ordering new supplies), one might expect improvement to begin somewhat after 1997. The lagged achievement trend in reading suggests that there may be subject-specific differences in the production function, making it more difficult to immediately increase reading performance.

To provide a better perspective of student achievement trends throughout the 1990s, Figure 2 shows the unadjusted achievement scores from 1990 to 2000.[10] A new form of the exam was introduced in 1993, which most likely accounts for the achievement jump that year. Otherwise the trends in Figures 1 and 2 tell a similar story—little change in achievement during

[10] It is not possible to replicate Figure 1 exactly because demographic information and prior achievement scores are not available for the earliest cohorts. In this figure, scores are standardized on 1990 achievement.

the early to mid-1990s, following by substantial achievement gains after the introduction of high-stakes testing.

To control for unobserved, time-varying factors at the state 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 Cleveland, 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.

Table 2 shows the OLS regression results that correspond to Figures 1 to 3. 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 since 1997, for example, one

would not want to include their fourth or fifth grade scores in the estimation.[11] 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.

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" as well as the fact that students and teachers may have become more efficient at responding to the policy over time (although 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 they may be most able to influence their own learning.[12] Table 3 shows the estimates reflecting the comparison between Chicago and other mid-western 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.

[11] 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.

[12] 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.

To provide a sense of the magnitude of these effects, one might consider the effect of the well-known Tennessee STAR experiments, in which students were randomly assigned to regular-size (22-26 students) or small-size (13-17 students) classrooms. The analysis of STAR found that attending a small class increased student achievement by roughly 0.15 to 0.25 standard deviations, with noticeably larger effects (0.25 to 0.35 standard deviations) for minority students (Krueger 1999, Nye et. al. 1999, Finn and Achilles 1999). It therefore appears that the Chicago accountability policy had an effect comparable to STAR in reading, and perhaps even larger in mathematics.[13]

5.2 The Heterogeneity of Effects Across Student and School Risk Level

The structure of the Chicago accountability policy suggests that it may generate larger effects for certain 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 20th percentile (on the national ability distribution) in reading and math; (2) the student margin for probation—to count toward the school's aggregate accountability measure, students needed to score above the 50th percentile nationally in reading; and (3) the

[13] 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 (2001a) 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 (2001a) 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.

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., 5^{th} grade scores for the 8^{th} grade cohorts).[14] Prior school achievement is based on the percent of students in the school in 1995 that met national norms on the reading exam.[15] The sample includes first-time students whose scores were included for reporting purposes. The latest

[14] 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.
[15] The results are robust to classifying school risk on the basis of achievement in other pre-policy years.

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 were over 40 percent of students had been meeting national norms. This is consistent with the fact that 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 10th-50th percentile in the past fared better than their classmates who had either scored below the 10th percentile, or above the 50th percentile. This is consistent with the incentives imposed on at-risk students by the policy to end social promotion. Moreover, this 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 math, or greater divisibility in the production of reading achievement. However, it is also important to note that these differential effects by student prior ability are considerably smaller than the differential effects by prior school ability. This suggests

that the response to the accountability policy took place at the school level, rather than the individual student level.

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. Thus, by examining the differential gains across subject and grade, it may be possible to separate the effect of the student and school-based accountability policy. Unfortunately, there are several difficulties in separately identifying these effects. 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. In addition, many of the 1998 fourth and seventh graders will have attended summer school the previous year. For this reason, this section focuses predominantly on results from the 1997 cohort.

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 these specifications, we have controlled for prior achievement in all three previous years (*t-1*,*t-2* and *t-3*) in order to make the cross-grade results as comparable as possible.[16] It appears that there is virtually no difference in the achievement effects across the two groups, in either 1997 or 1998.[17] One explanation for this finding is that the school probation policy was driving the overall achievement results.

[16] Because we are focusing on the 1993-1997 cohorts, this full set of prior achievement controls is not endogenous to the policy, unlike when we use later cohorts.
[17] The results are similar across the ability distribution. Tables available from the author upon request.

Alternatively, students in grades four, five and seven may have incorrectly believed that they were subject to the promotional requirements. Student interviews provide some evidence for this confusion, possibly because teachers in these grades emphasized the promotional criteria to motivate students.[18] 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.

5.4 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 the 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. A difficulty in comparing achievement across subjects in Chicago is not only that science and social studies exams are not given in certain grades, but also that the grades in which the subjects are given has changed over time. For this reason, we are forced to limit our analysis here to grades four and eight, from 1995 to 1998.[19]

Table 6 shows the impact of the accountability policy across subjects. We see that achievement gains in math and reading were roughly two to four times larger than gains in

[18] For more information on qualitative studies of the accountability policy in Chicago, see Engel and Roderick (2001).

[19] 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.

science and social studies, although it is important to note that science and social studies scores also increased under HST. Moreover, the distribution of effects is somewhat different for low versus high-stakes subjects. As we noted earlier, in math and reading, students in low-achieving schools experienced greater gains although, 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).[20]

5.5 Other Student Outcomes: Test-Taking, Special Education and Off-Grade Retention

While the accountability policies in Chicago are designed to increase student achievement, they also create incentives for students and teachers that may change test-taking patterns.[21] A certain number of students do not take the ITBS each year, either because they are absent on the exam day or because they are exempt from testing due to placement in certain 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

[20] This also suggests that if one used science and social studies achievement as a counterfactual in the estimation of the HST effects, then one might find significantly larger effects for low-achieving students. Because of the limited data on science and social studies scores, however, I did not construct any formal difference-in-difference estimates using the variation across subject.

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.[22] 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 4 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 appears that the trend has become 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.

---

[21] 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.

[22] 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

Table 7 shows the corresponding Probit estimates for special education placement. The sample is limited to the 1994-1998 cohorts because estimates for the later cohorts may be confounded by earlier grade retention.[23] 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 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 lower relative to school average achievement level and were thus more obvious candidates for evaluation. Here we see that the highest risk students, conditional on their prior achievement level,[24] 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

---

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.

[23] 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.

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

percent among moderate-achieving schools and no increase among the highest performing schools. This is consistent with the incentives provided by the policy.

Another way for teachers to shield low-achieving students from the accountability mandates is to preemptively retain them. 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. This suggests that even in grades not directly affected by the promotional policy retention rates may have increased under high-stakes testing.[25] 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 new students entering these grades become more likely to have been retained in earlier grades.[26] Figure 5 shows this exact pattern. Prior to the accountability policy, the retention rate was roughly 4 to 5 percent in first 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, which may have been in anticipation of the new standards the students would face in 1997. In most grades, the rates peaked in 1997 and then declined somewhat. The first grade retention rate continued to increase over time, most likely because it is the first year in which many students would be retained (in contrast to other grades, in which teachers would be taking into consideration prior retentions in deciding whether to hold a student back).

Table 8 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

[25] 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.

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.

5.6 Sensitivity Analysis

To test the sensitivity of the findings presented in the previous sections, Table 9 presents comparable estimates for a variety of different specifications and samples. Since the results are comparable across cohort, for simplicity I only present the results for the 1998 cohort. Column 1 shows the baseline estimates. Column 2 shows that including students with missing outcome data does not change the results. Column 3 shows that the results are robust to not including a pre-existing achievement trend. Columns 4-6 show that the results do not change if various groups are excluded from the sample or if school fixed effects are included. Column 7 shows that the results are not sensitive to the inclusion of prior achievement measures, suggesting that the composition of students did not change substantially over this period.

[26] 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.

6. The Effect of the Accountability Policy on Low-Stakes versus High-Stakes Exams

6.1 The Influence of High-Stakes Testing on IGAP Scores

Under the Chicago accountability policy, student promotion and school probation are based entirely on student achievement on the Iowa Test of Basic Skills (ITBS), a standardized exam that has been administered by the district for many years. However, students in Chicago also take a state-administered achievement exam known as the Illinois Goals Assessment Program (IGAP). In fact, prior to 1996, the IGAP was the higher-stakes exam for Chicago (and still is for many districts in Illinois), although the stakes were largely indirect and relatively minor compared with the consequences associated with the new accountability policy. 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." Throughout this period, the ChiPS placed little if any emphasis on the ITBS. In 1996, the situation changed dramatically. The ChiPS placed large incentives on the ITBS results while the IGAP incentives remained the same.

The administration of multiple exams in Chicago allows us to more carefully examine the effect of the accountability policy. Figure 6 shows IGAP achievement trends in Chicago relative to other urban districts in Illinois.[27] The data for this analysis is drawn from school "report

[27] 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

cards" compiled by ISBE, which provide average IGAP scores by grade and subject as well as background information on schools and districts. 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 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. Chicago appears to have narrowed the achievement gap during the 1990s. 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. However, achievement scores in grade eight, particularly in reading, did show more of a break in trend in 1996.[28]

Table 10 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.

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.

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

As on the ITBS, low-achieving schools made larger gains on the IGAP than high-achieving schools. Table 11 shows estimates for grades three, six and eight together by school 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 greater than 40% meeting norms). In the lowest-achieving schools, we see that 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 to control for other unobserved state factors. 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 Explaining the Differential Gains on the ITBS and the IGAP

It appears that the accountability policy led to large increases in ITBS scores, but had zero or perhaps even negative effect on IGAP scores. For both exams, students in the lowest performing schools made larger gains relative to their peers in higher performing schools. However, even in low-performing schools, IGAP scores did not deviate significantly from pre-existing trends. These findings are consistent with several earlier studies of test-based accountability (Linn and Graue 1990, Shepard 1990, Koretz et. al. 1991, Koretz and Barron 1998, Stecher and Barron 1998, Klein et. al. 2000).

There are several possible explanations for the differential trends. First, the accountability policy may have led teachers to focus on topics tested in the ITBS and de-emphasize those skills measured by the IGAP. For this type of substitution to result in differential achievement patterns, the exams must differ at least partially in terms of content or format. If this were the case, however, it is still not clear how to interpret the finding. To the extent that the exams truly measure different topics or skills, and the ITBS gains are an accurate indication of increasing knowledge in these areas, then one should think of the differential achievement patterns in terms of a tradeoff between different skill sets. On the other hand, if the exams measure similar underlying concepts and differ primarily in the way in which the concepts are assessed, then one might view smaller gains on the IGAP as an indication that the accountability policy has simply increased test-taking skills and not truly enhanced student learning. Another explanation is that students simply exerted greater effort on the ITBS (relative to the IGAP) following the introduction of the accountability program. Even if the accountability policy increased student effort on standardized testing in general, it may have induced a larger relative on the ITBS than the IGAP. In this view, the differences in testing conditions prevent one from drawing any inferences about differential learning rates. One story consistent with the observed trends involves a combination of effort and test-preparation/curriculum alignment. If the accountability policy increased student effort on standardized testing in general, but also led to a shift away from IGAP-specific material, one might observe a continuing upward trajectory on the IGAP at the same time as a sharp increase on the ITBS.[29]

[29] Another explanation involves teacher cheating. While Jacob and Levitt (2002) found that instances of classroom cheating increased substantially following the introduction of high-stakes testing, they estimate that cheating increases could only explain an extremely small part of the ITBS gains since 1996-97.

Unfortunately, because many of these factors such as effort and test preparation are not directly observable, it is difficult to attribute a specific portion of the gains to a particular cause. Instead, we seek to provide some evidence regarding the potential influence of test preparation/curriculum alignment and effort.

6.2.1 The Role of Test Preparation

One explanation for the differential achievement trends involves test-specific preparation. If the ITBS and IGAP measure different topics or skills and teachers have aligned their curriculum to the ITBS in response to the accountability policy, then we might see disproportionately large increases on the ITBS.[30] As we can see in Table 12, while the two exams have the same general format, the IGAP appears to place somewhat greater emphasis on critical thinking and problem-solving skills. For example, 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 its long passages, multiple correct answers and questions comparing passages, the IGAP reading exam appears to be more difficult and more heavily weighted toward critical thinking skills than the ITBS exam.

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 the ITBS items that are (a) easy to teach to and/or (b) relatively more common on the ITBS than the IGAP. Table 13 presents OLS estimates of the relationship between high-stakes testing and ITBS math achievement by item type. The sample is limited to the 1996 and 1998 cohorts, which took ITBS Form L. The dependent variable is the proportion of students who answered the item correctly in

the particular year. We see that students made the largest gains on items involving computation and number concepts and made the smallest gains on problems involving estimation, data analysis and problem-solving skills. For example, the estimates in column 1 indicate that students in 1998 were 2.3 percentage points more likely to correctly answer questions involving critical thinking (i.e., data analysis, estimation and problem-solving) and 4.2 percentage points more likely to correctly answer basic skill questions (i.e., number concepts and math computation). Column 2 shows that the largest relative increase came on math computation questions. While not conclusive, this suggests that teachers may have shifted the focus of instruction to better match the ITBS content.

6.2.2 The Role of Effort

An alternative explanation involves student effort. If the consequences associated with ITBS performance led students to concentrate harder during the exam or caused teachers to ensure optimal testing conditions for the exam, student achievement on the ITBS may have exceeded IGAP achievement even if the content of the exams were identical. One indication of effort involves test completion. Prior to the introduction of high-stakes testing, roughly 20 to 30 percent of students left items blank on the ITBS exams despite the fact that there was no penalty for guessing. 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 correctly *increased* by roughly 9 percent at the same time that the percent blank has declined by 36 percent, suggesting that the higher

[30] Tepper (2002) analyzed teacher surveys in 1994, 1997 and 1999 and found that teacher-reported test preparation

completion rates were not due entirely to guessing. 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 20 percent of the observed ITBS gains.

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. Table 14 presents OLS estimates of the relationship between item position and achievement gains from 1994 to 1998. The analysis is limited to the reading exam because the math exam is divided into several sections, so that item position is highly correlated with item type. Conditional on item difficulty, student performance on the last 20 percent of items increased 2.7 percentage points more than performance on the first 20 percent of items. Based on these results, it appears that nearly all of the improvement on the ITBS reading exam came at the end of the test. This suggests that effort may have played a significant role in the ITBS gains seen under high-stakes testing.[31]

7. Conclusions

I find that the introduction of test-based accountability policy in Chicago generated a substantial increase in student achievement. Math and reading scores increased sharply following the introduction of the accountability policy, suggesting that teachers and students

---

and curriculum alignment increased following the introduction of the accountability policy.

responded quite strongly to the incentives. Moreover, students in low-achieving schools experienced larger gains than their peers in other schools, and low to moderate achieving students showed larger gains than higher ability students, consistent with the incentives generated by the accountability policy.

However, I also find that teachers and administrators responded strategically to the incentives along a variety of other dimensions, some of which might have undesirable consequences. Following the introduction of high-stakes testing, (i) math and reading scores increased relative to scores on low-stakes exams such as science and social studies; (ii) there was a substantial increase in the proportion of students in special education; and (iii) retention rates increased substantially in grades not directly affected by the student promotion policy. In addition, I find that student performance on a low-stakes exam decreased relative to performance on the high-stakes exam. This appears at least partly due to test preparation and curriculum alignment directed toward the high-stakes exam along with greater student effort on the high-stakes exam.

These results suggest that policy-makers must approach the high-stakes testing with caution. Without appropriate safeguards, the incentives created by a test-based accountability system might lead to undesirable consequences for students. In particular, state education agencies must consider these issues in implementing the new federal legislation.

[31] An alternative explanation is that students prior to 1997 did not finish the exam because they could not answer the items quickly enough, but that after accountability, student learning increasing and thus allowed them to complete the exam faster and correctly answer more items at the end of the exam.

References

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.

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.

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).

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).

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.

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 (2001a). "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." Working Paper.

Jacob, B. A. and L. Lefgren (2001b). "The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from Reform Efforts in Chicago." Working Paper.

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

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.

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 (2000). The Grasshopper and the Ant: Motivational Responses of Low-Achieving Students to High-Stakes Testing. American Educational Research Association, New Orleans.

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.

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	0.201 (0.040)	0.276 (0.043)
1999 Cohort	0.221 (0.034)	0.213 (0.035)
1998 Cohort	0.179 (0.022)	0.234 (0.024)
1997 Cohort	0.058 (0.020)	-0.045 (0.021)
6th Grade		
2000 Cohort	0.178 (0.022)	0.322 (0.027)
1999 Cohort	0.147 (0.018)	0.176 (0.022)
1998 Cohort	0.194 (0.013)	0.240 (0.015)
1997 Cohort	0.102 (0.011)	0.105 (0.013)
8th Grade		
2000 Cohort	0.251 (0.023)	0.436 (0.025)
1999 Cohort	0.224 (0.020)	0.447 (0.028)
1998 Cohort	0.179 (0.014)	0.276 (0.015)
1997 Cohort	0.113 (0.012)	0.289 (0.013)
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 Cleveland, 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

	1997		1998		
Dependent Variable	Student + School Incentive (Grades 3, 6 & 8)	School Incentive (Grades 4, 5 & 7)	Student + School Incentive (Grades 3, 6 & 8)	School Incentive (Grades 4, 5 & 7)	School Incentive (Grade 5)
Math Score	0.064 (0.007)	0.076 (0.006)	0.122 (0.008)	0.120 (0.007)	0.036 (0.012)
Reading Score	0.097 (0.008)	0.105 (0.007)	0.139 (0.008)	0.185 (0.007)	0.072 (0.012)

Notes: The sample includes first-time students who were tested and whose scores were included in reporting. The estimates shown are the coefficients on indicators for cohorts that experienced high-stakes testing (1997 or 1998). Robust standard errors that account for the correlation of errors within schools are shown in parenthesis.

Table 6: 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 * (< 10^{th} percentile)	-0.030 (0.023)	0.014 (0.022)	-0.081 (0.022)	-0.069 (0.022)
HS * (10-25^{th} percentile)	-0.040 (0.017)	0.018 (0.015)	-0.065 (0.017)	-0.058 (0.017)
HS* (26-50^{th} percentile)	-0.028 (0.014)	0.014 (0.013)	-0.032 (0.015)	-0.029 (0.015)
HS * (< 20% students scored above the 50^{th} percentile)	0.083 (0.022)	0.097 (0.020)	0.035 (0.022)	0.030 (0.023)
HS* (20-40% students scored above the 50^{th} 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 7: 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 8: 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 9: Sensitivity Analysis

	Specification						
	Baseline	Missing outcome data imputed at 25th percentile in school	No pre-existing trend	Excluding retained students	Excluding retained students and students whose scores were not reported	Including school FE	No Prior Achievement Controls
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Reading							
3rd Grade	0.199 (0.018)	0.200 (0.018)	0.229 (0.012)	0.215 (0.018)	0.170 (0.019)	0.190 (0.017)	--
6th Grade	0.193 (0.012)	0.195 (0.012)	0.195 (0.009)	0.194 (0.013)	0.200 (0.013)	0.182 (0.012)	0.130 (0.015)
8th Grade	0.182 (0.014)	0.185 (0.014)	0.160 (0.009)	0.182 (0.014)	0.187 (0.014)	0.178 (0.014)	0.123 (0.017)
Math							
3rd Grade	0.249 (0.020)	0.255 (0.020)	0.301 (0.013)	0.248 (0.020)	0.209 (0.021)	0.243 (0.019)	--
6th Grade	0.239 (0.014)	0.240 (0.015)	0.224 (0.010)	0.227 (0.015)	0.227 (0.016)	0.232 (0.014)	0.192 (0.016)
8th Grade	0.282 (0.015)	0.284 (0.015)	0.232 (0.010)	0.277 (0.015)	0.282 (0.015)	0.277 (0.015)	0.233 (0.018)

Notes: Results based on the 1998 cohort.

Table 10: The Impact of Test-Based Accountability on Low-Stakes Achievement Test Scores

	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 HST years)						
(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 HST years)						
(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 11: 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 10. The regressions are weighted by the inverse square root of the number of students enrolled in the school.

Table 12: A Comparison of Eighth Grade ITBS and IGAP Exams

	Math		Reading	
	ITBS	**IGAP**	**ITBS**	**IGAP**
Structure	• 135 multiple-choice questions • 4 possible answers • No penalty for wrong answers • Five sessions of 20-45 minutes each	• 70 multiple-choice questions • 5 possible answers • No penalty for wrong answers • Two 40 minute sessions	• 7 passages followed by 3-10 multiple-choice questions • 49 total questions	• 2 passages followed by 18 multiple-choice questions • 4 questions that ask the student to compare the two passages • 40 questions total
Content	• Computation (43) • Number Concepts (32) • Estimation (24) • Problem-Solving (20) • Data Analysis (16)	• Computation (10) • Ratios & Percentages (10) • Measurement (10) • Algebra (10) • Geometry (10) • Data Analysis (10) • Estimation (10)	• 2 narrative passages • 4 expository passages • 1 poetry passage	• 1 narrative passage • 1 expository passage
Format	• Computation problems do not have words. • Data Interpretation section consists of a graph or figure followed by several questions.	• All questions are written as word problems, including the computation problems. • One question per graph or figure.	• One correct answer per question.	• Multiple correct answers per question.

Notes: Information on the ITBS is taken from the Form L exam. Information on the IGAP is based on practice books.

Table 13: OLS Estimates of the Relationship between Item Type and Achievement Gain on ITBS Math Exam from 1996 to 1998

	Dependent Variable = Proportion of Students Answering the Item Correctly on the ITBS Math Exam	
Independent Variables	**Model 1**	**Model 2**
1998 Cohort	.023 (.011)	.015 (.012)
Basic Skills * 1998	.019 (.006)	
Number Concepts *1998		.019 (.010)
Estimation *1998		.006 (.010)
Data Analysis *1998		.008 (.011)
Math Computation *1998		.029 (.009)
25-35% answered item correctly prior to high-stakes testing*1998	.015 (.012)	.013 (.012)
35-45% answered item correctly prior to high-stakes testing*1998	.019 (.012)	.019 (.012)
45-55% answered item correctly prior to high-stakes testing*1998	.018 (.012)	.016 (.012)
55-65% answered item correctly prior to high-stakes testing*1998	.015 (.013)	.012 (.013)
65-75% answered item correctly prior to high-stakes testing*1998	.010 (.014)	.009 (.014)
75-85% answered item correctly prior to high-stakes testing*1998	-.002 (.014)	-.005 (.014)
85-100% answered item correctly prior to high-stakes testing*1998	.003 (.024)	-.003 (.024)
Number of Observations	692	692
R-Squared	.956	.957

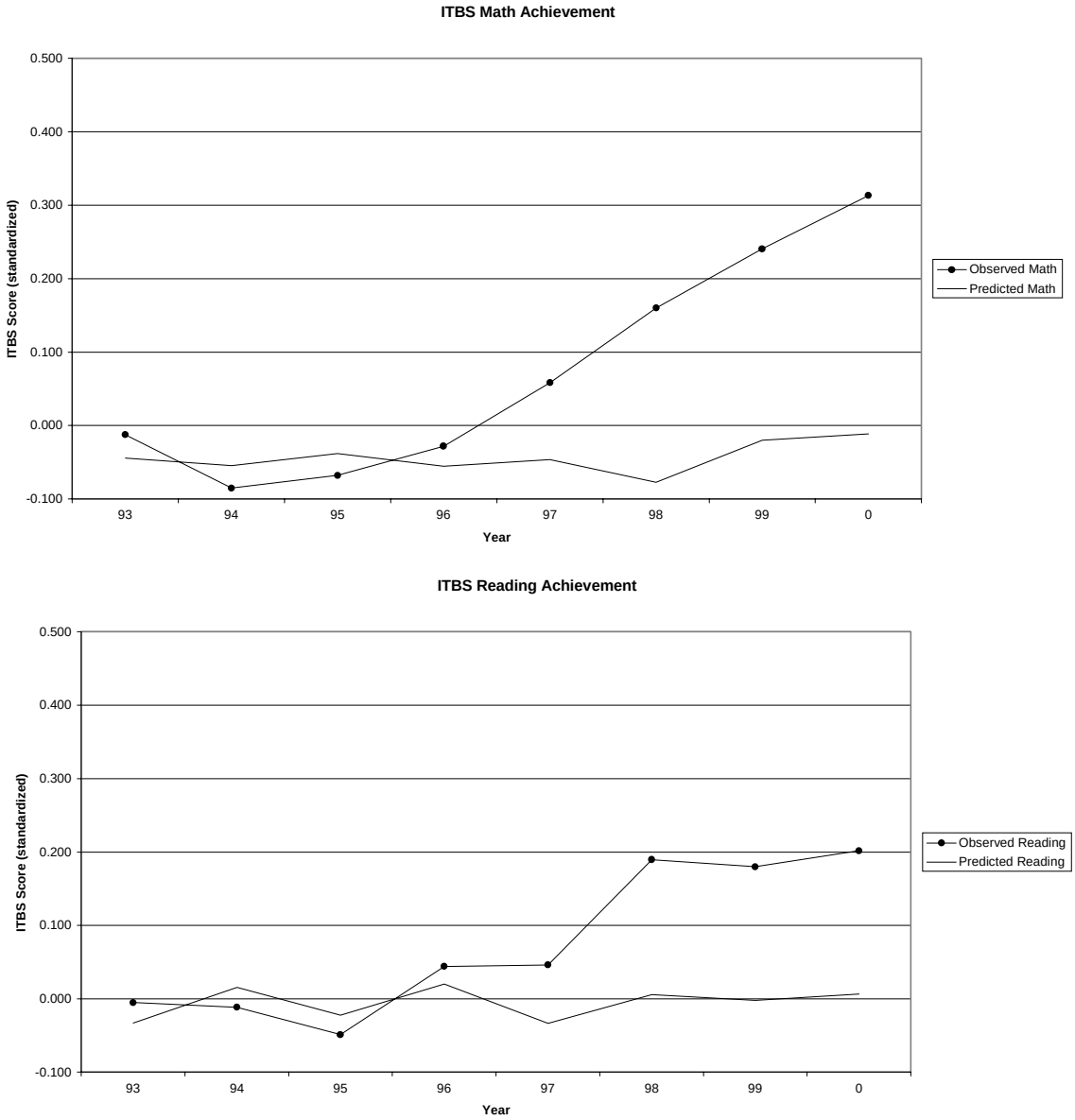
Notes: The sample consists of all tested and included students in 1996 and 1998. The units of observation are item*year proportions, reflecting the proportion of students answering the item correctly in that year. Model 1 categorizes items as either basic-skills or critical thinking, using the latter group as the omitted category. Model 2 categorizes items into five types, using problem-solving as the omitted category.

Table 14: OLS Estimates of the Relationship between Item Position and Achievement Gain on the ITBS Reading Exam from 1994 to 1998

	Dependent Variable = Proportion of Students Answering the Item Correctly on the ITBS Reading Exam
	Total
Intercept	.004 (.021)
2nd Quintile of the Exam	.002 (.014)
3rd Quintile of the Exam	.013 (.015)
4th Quintile of the Exam	.017 (.015)
5th Quintile of the Exam	.027 (.017)
25-35% answered item correctly prior to high-stakes testing	.025 (.020)
35-45% answered item correctly prior to high-stakes testing	.036 (.019)
45-55% answered item correctly prior to high-stakes testing	.049 (.019)
55-65% answered item correctly prior to high-stakes testing	.046 (.021)
65-75% answered item correctly prior to high-stakes testing	.051 (.025)
75-100% answered item correctly prior to high-stakes testing	.043 (.030)
Number of Observations	258
R-Squared	.95

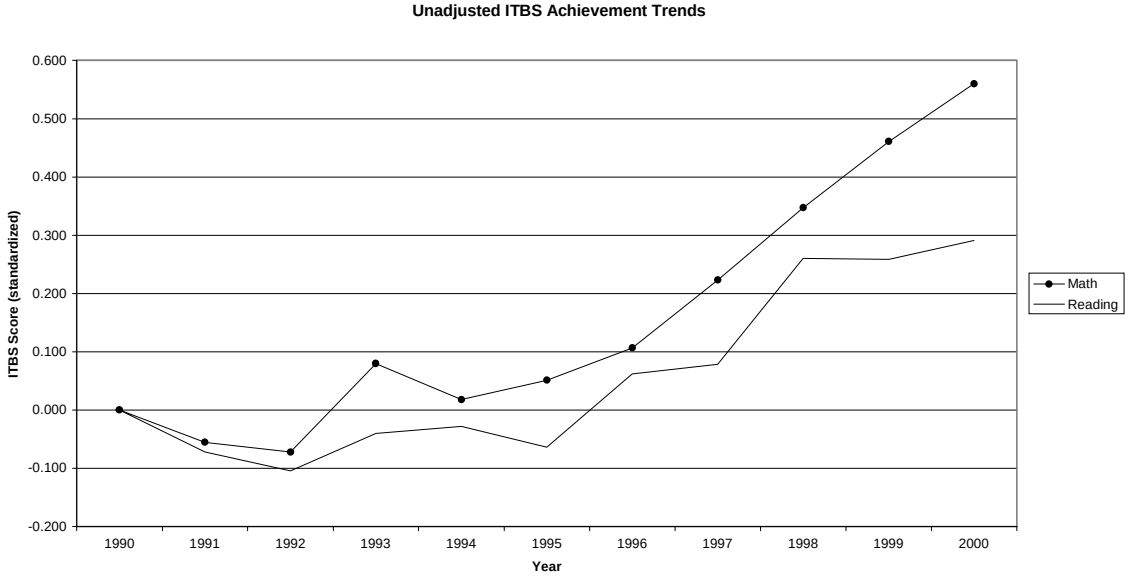
Notes: The sample consists of all tested and included students in 1994 and 1998. The units of observation are item*year proportions, reflecting the proportion of students answering the item correctly in that year. The omitted category is the first quintile of the exam.

Figure 1: ITBS Achievement Trends in Chicago, 1993-2000



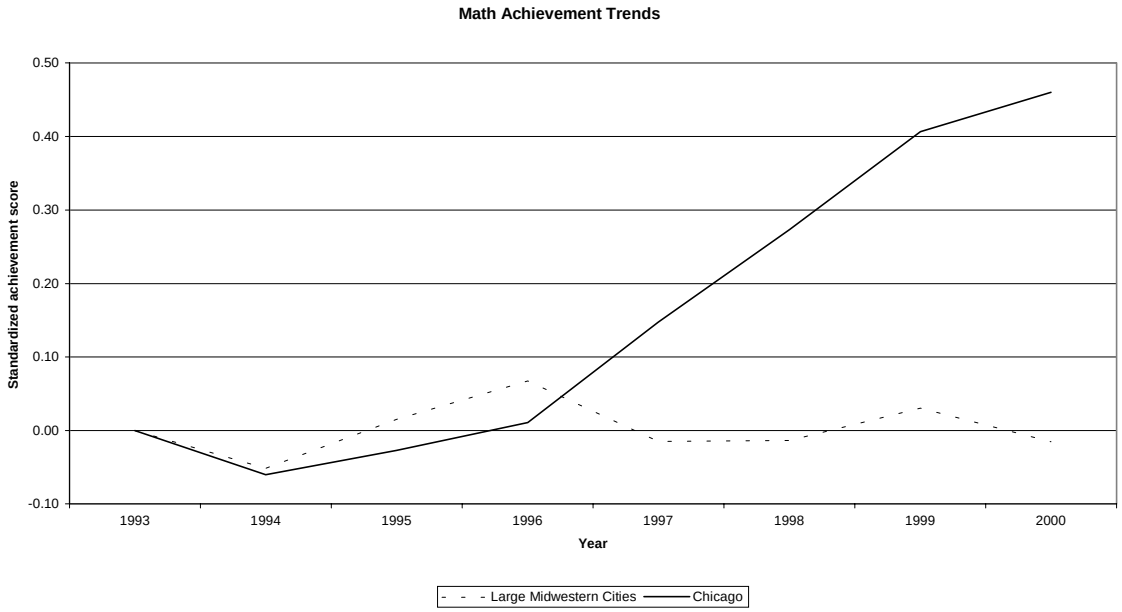
Notes: The sample includes 3rd, 6th and 8th grade students from 1993 to 2000, excluding retainees and students whose scores were not reported. Scores are standardized separately for each grade using the 1993 student-level mean and standard deviation. The predicted scores are derived from an OLS regression 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.

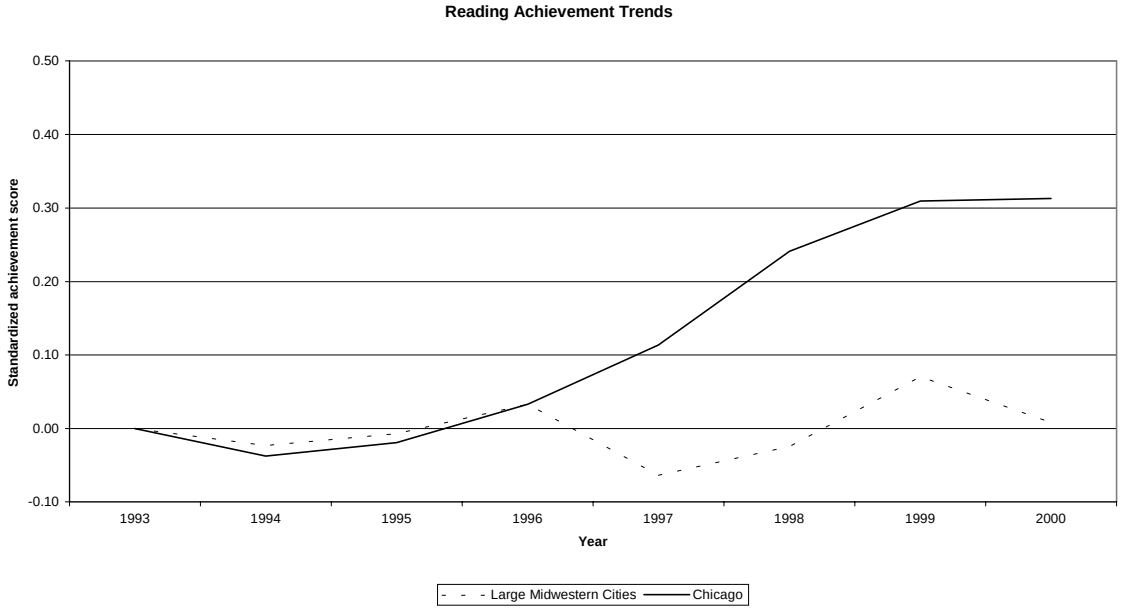
Figure 2: Unadjusted ITBS Achievement Trends in Chicago, 1990-2000



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.

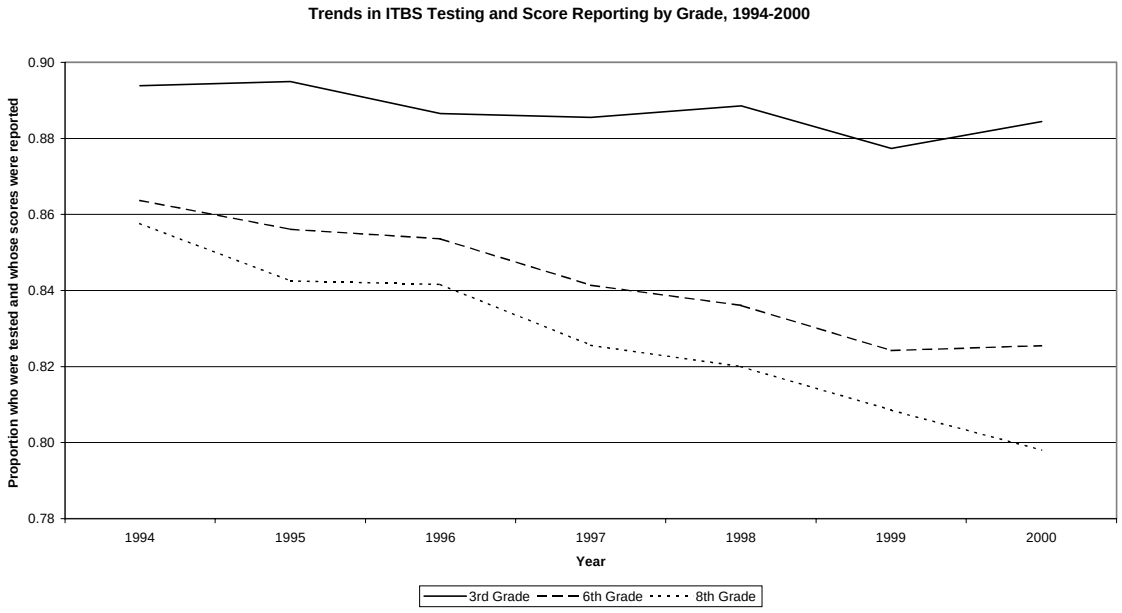
Figure 3: Achievement Trends Across the Nation During the 1990s

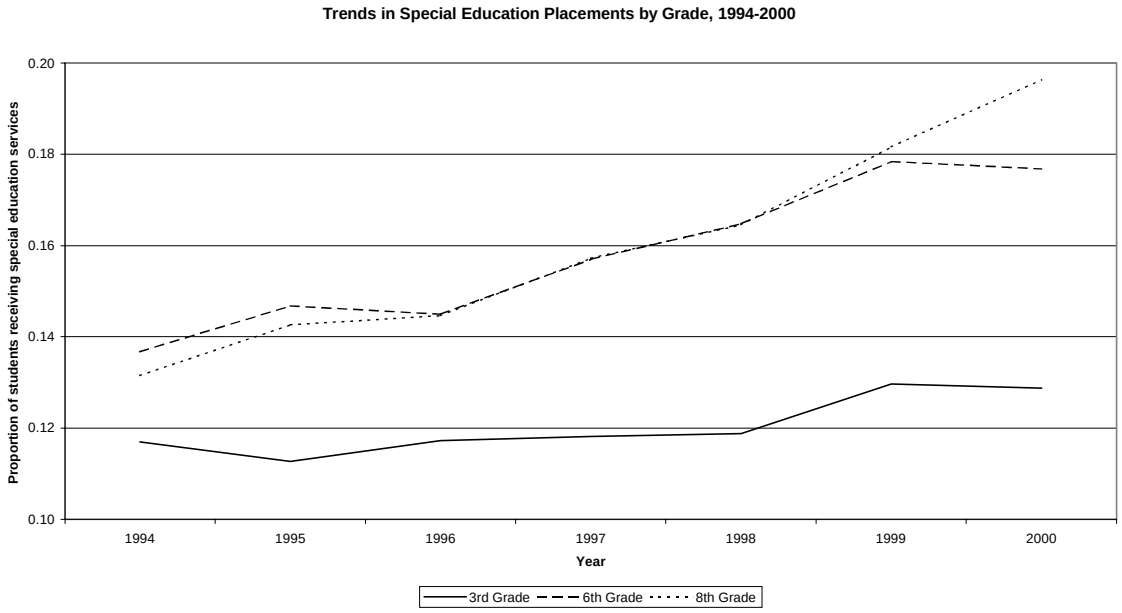




Notes: The achievement series for large Midwestern cities includes data for all tested elementary grades in Cleveland, Cincinnati, Gary, Indianapolis, St. Louis and Milwaukee.

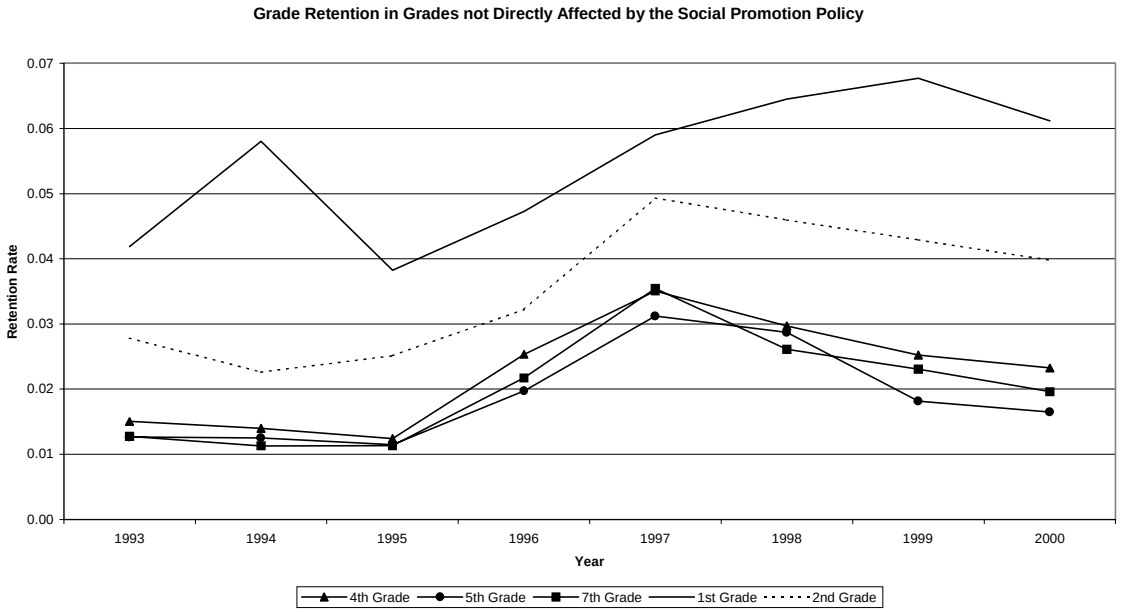
Figure 4: Trends in Testing and Special Education Placements





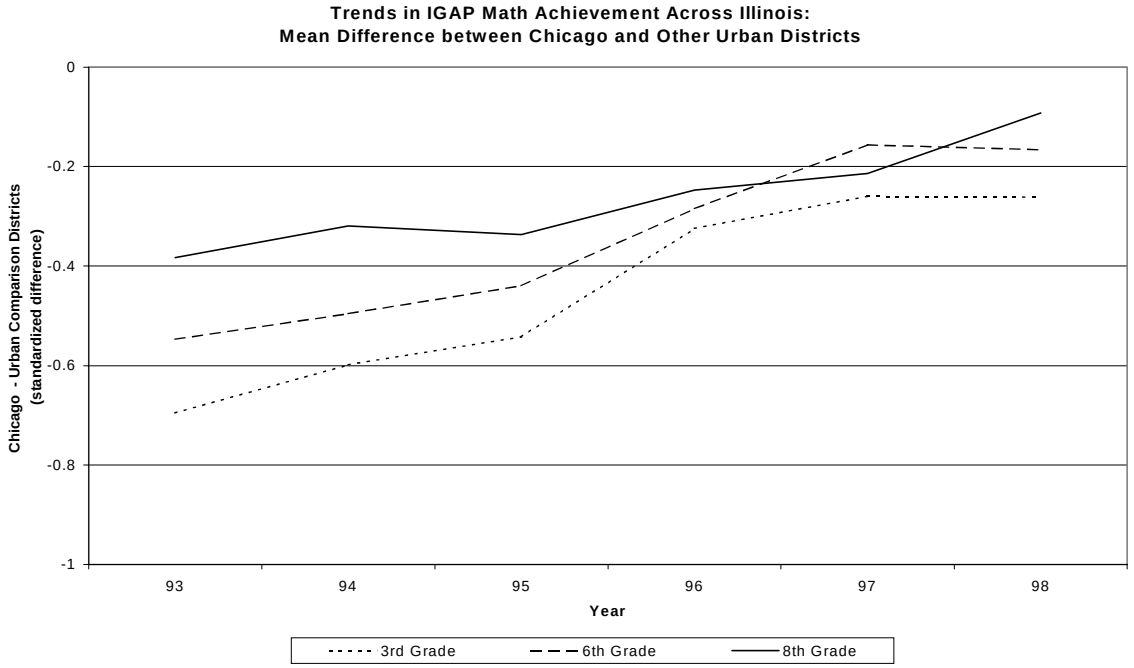
Notes: The sample includes only first-time, non-bilingual students.

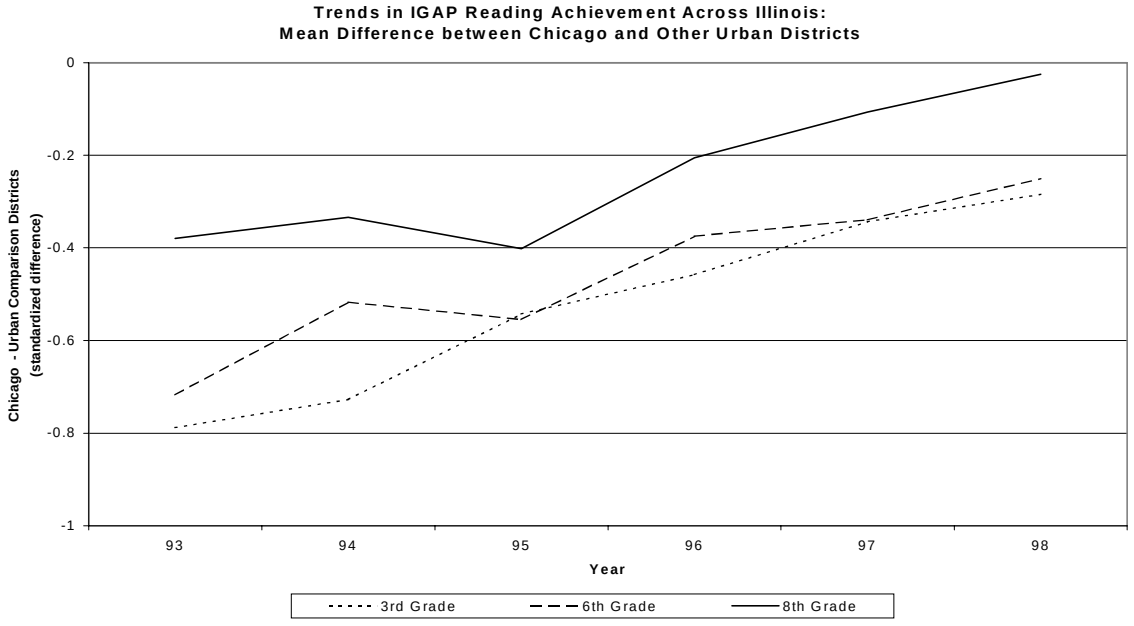
Figure 5: Trends in Grade Retention



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

Figure 6: 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.