

AS THE FIRST LARGE URBAN SCHOOL DISTRICT TO INTRODUCE a comprehensive accountability system, Chicago provides an exceptional case study of the effects of high-stakes testing—a reform strategy that will become omnipresent as the No Child Left Behind Act is implemented nationwide. One of the most serious criticisms of high-stakes testing is that it leads to “inflated” test scores that do not truly reflect students’ knowledge or skills and therefore cannot be generalized to other tests. This article summarizes my research on whether the Chicago accountability system produced “real” gains in student achievement.

The first step in Chicago’s accountability effort was to end the practice of “social promotion,” whereby students were advanced to the next grade regardless of achievement level. Under the new policy, students in the 3rd, 6th, and 8th grades were required to meet minimum standards in reading and mathematics on the Iowa Test of Basic Skills (ITBS) in order to step up to the next grade. Students who didn’t meet the standard were required to attend a six-week summer-school program, after which they took the exams again. Those who passed were able to move on to the next grade. Students who again failed to meet the standard were required to repeat the grade, with the exception of 15-year-olds who attended newly created “transition” centers. (Many students in special education and bilingual programs were exempt from these requirements.) In the fall of 1997, roughly 20 percent of

ILLUSTRATION BY NOAH WOODS

HIGH STAKES by BRIAN JACOB in CHICAGO

Did Chicago’s rising test scores reflect genuine academic improvement?

1934

A	B	C	D	E	F	G	H	I	J
K	L	M	N	O	P	Q	R	S	T
U	V	W	X	Y	Z				
1	2	3	4	5	6	7	8	9	0
!	,"	;	:	?	~	^	&	*	^

s LES CLASSIQUES

Maestro (i
Saint

1959

$$\begin{array}{r} 2 \times 5 \\ \hline 7^2 \end{array}$$

$$\begin{array}{r} 730^5 \\ \hline 114 \end{array}$$

50178

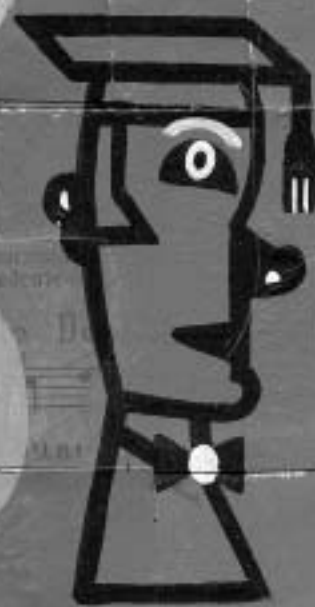
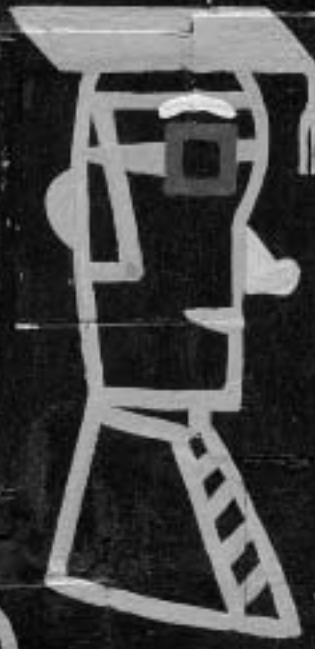
094

5 23

LE

EDEN, m. G
y bor al ergo, p

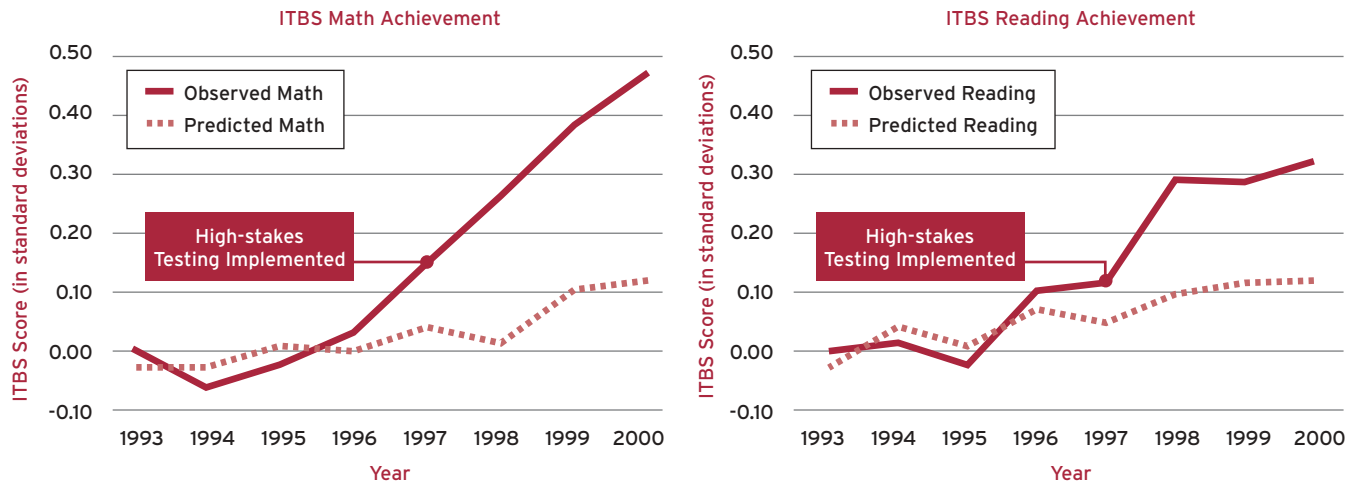
Do sosten



674

Exceeding Expectations (Figure 1)

After high-stakes testing was implemented, Chicago's scores on the Iowa Test of Basic Skills increased at a faster-than-predicted pace in both math and reading



Note: The sample includes 3rd, 6th, and 8th grade students from 1993 to 2000, excluding students who were retained in grade. Predicted scores account for changes in student composition and prior achievement levels, and for achievement trends before 1997.

SOURCE: Author

Chicago's 3rd graders and 10 to 15 percent of 6th and 8th graders were held back.

Meanwhile, Chicago also instituted an "academic probation" program designed to hold teachers and schools accountable for student achievement. Schools in which fewer than 15 percent of students scored at or above national norms on the ITBS reading exam were placed on probation. If they did not exhibit sufficient improvement, these schools could be reconstituted, with teachers and school administrators dismissed or reassigned. In the 1996-'97 school year, 71 elementary schools were placed on academic probation. While only recently has Chicago actually reconstituted several schools, as early as 1997 teachers and administrators in probationary schools reported being extremely worried about their job security, and staff in other schools reported a strong desire to avoid probation.

High Stakes and Test Scores

Scores on the ITBS increased substantially in Chicago in the second half of the 1990s. However, many factors besides

the accountability policies may have influenced the achievement trends in Chicago. For instance, the population of students may have changed during the period in which high-stakes testing was implemented. An influx of recent immigrants during the mid- to late 1990s may depress the city's test scores, whereas they would be likely to rise with the return of middle-class students to the city. Similarly, policy changes at the state or national level, such as the efforts to reduce class sizes or mandate higher-quality teachers, if effective, would likely lead one to overestimate the impact of Chicago's policies.

The rich set of longitudinal, student-level data available for Chicago allowed me to overcome many of these concerns. I was able to adjust for observable changes in student composition, such as the district's racial and socioeconomic makeup and its students' prior achievement. Moreover, because achievement data were available back to 1990, six years prior to the introduction of the accountability policies, I was able to account for preexisting achievement trends within Chicago. Using this information, I looked for a sharp increase in

achievement (a break in trend) following the introduction of high-stakes testing as evidence of a policy effect. Comparing achievement trends in Chicago with those in other urban districts in Illinois as well as in large midwestern cities outside Illinois enabled me to address the concern about actions at the state and federal level that might have influenced achievement.

The sample consisted of students who were in the 3rd, 6th, and 8th grades from 1993 to 2000. The new policy on social promotion caused a large number of low-performing students in these grades to be retained, substantially changing the student composition in these and subsequent grades beginning in the 1997-'98 school year. For this reason I limited the sample to students who were in these three grades for the first time in their school career. Moreover, the results presented here are based on only those students who were tested and whose scores were included by the district for official reporting purposes. (Each year roughly 10 percent of students were not tested, and an additional 10 to 15 percent had scores that were not reported because of a special education

or bilingual placement.) Analyses using a sample of all students who were tested yielded similar results. While special education placement rates appeared to increase following the introduction of the accountability policy in Chicago, this alone can explain only a small fraction of the observed achievement gains.

Figure 1 plots the predicted versus the observed achievement scores for successive cohorts of Chicago students from 1993 to 2000. Predicted scores were obtained from a regression analysis that accounted for changes in student composition and prior achievement levels as well as overall trends in achievement before the introduction of the accountability program. The figure indicates that neither observable changes in student composition nor preexisting achievement trends in Chicago explain the substantial improvement in student performance since 1997. The trends predicted that achievement in math would decrease or remain flat after 1996. In practice, however, achievement slipped somewhat from 1993 to 1996, but increased sharply after 1996. By 2000, math scores were roughly 0.3 standard deviations higher than predicted, an improvement about one quarter the size of the difference in math performance between Chicago students in consecutive grades in 1995. A similar pattern was apparent in reading. Predicted and observed test scores were relatively flat from 1993 to 1996. In 1997 the gap between observed and predicted scores appeared to widen, and then grew substantially in 1998. By 2000 students were scoring roughly 0.2 standard deviations higher than predicted.

Still, it is possible that the achievement gains in Chicago simply reflected improvements in student performance in the state or nation. The economy was growing throughout the latter half of the 1990s, and there was a considerable emphasis on public education at the federal level. The achievement of a nationwide sample of 4th and 8th grade students with the same racial make-up as

Administrators in probationary schools reported being extremely worried about their job security.

Chicago students, as measured by the National Assessment of Educational Progress (NAEP), increased roughly 0.25 standard deviations in math during the 1990s, though there was no gain in reading.

However, a comparison with other urban districts in Illinois and the Midwest, such as Cincinnati, Gary, Indianapolis, Milwaukee, and St. Louis, none of which created a similar accountability system during this period, shows that Chicago's trend is unique (see Figure 2). Trends in Chicago and the other cities tracked one another remarkably well from 1993 to 1996, then began to diverge in 1997. Math and reading achievement in the comparison districts remained relatively constant from 1996 to 2000, while achievement levels in Chicago rose sharply over this period—by roughly 0.3 standard deviations in math and 0.2 standard deviations in reading.

Together, these results suggest that the accountability policy in Chicago led to a substantial increase in math and reading achievement. It appears that the effects were somewhat larger for math than for reading. This is consistent with a number of studies that show larger effects in math than in reading, presumably because reading achievement is more strongly influenced by family and other factors besides schooling. The effects were also somewhat larger for 8th grade students. This is consistent with the fact that 8th graders faced the largest incentives: they could not move to high school with their peers if they failed to meet the standards for promotion.

Checking on Inflation

The accountability policies that the Chicago school system put in place clearly led to an increase in scores on the ITBS. Nevertheless, critics of high-stakes testing wonder whether those increases reflect real gains in students' knowledge and skills—gains that ought to translate to students' performance in school and on other exams. When test scores are "inflated"—by, say, cheating or intense preparation that is geared to a specific exam—observed achievement gains are misleading because they do not reflect a more general mastery of the subject.

Researchers have found considerable evidence of test-score inflation throughout the country during the past two decades. In 1987, for example, John Jacob Cannell discovered what has become known as the "Lake Wobegon" effect—the fact that a disproportionate number of states and districts report being "above the national norm." More recently, researchers have demonstrated that Kentucky and Texas made substantially larger gains on state tests (the KIRIS and TAAS, respectively) than on the National Assessment of Educational Progress (NAEP). This is one reason why the No Child Left Behind legislation requires states to consider NAEP scores along with scores from state exams.

An approach commonly used to investigate score inflation is to compare student performance trends across exams. The notion is that if the gains on the high-stakes exam are not accompanied by gains on other achievement exams, then the gains may not be generalizable.

In Chicago, elementary students have traditionally taken two exams. The district has administered the ITBS, one of several standardized, multiple-choice exams used by districts across the country, to students in grades 3 to 8 for many years. Chicago's accountability sanctions were determined solely by student performance on the ITBS, making it the

high-stakes exam. At the same time, Chicago elementary students took another standardized, multiple-choice exam administered by the state, known as the Illinois Goals Assessment Program (IGAP). Before 1996 the IGAP was arguably the higher-stakes exam, even though there were no direct consequences for students or schools tied to the IGAP, since results from it appeared annually in local newspapers. After 1996 the IGAP clearly became the low-stakes exam for students and teachers in Chicago in comparison with the ITBS.

In 1993, Chicago students scored between 0.4 and 0.8 standard deviations below students in other urban districts on the IGAP. During the mid-1990s, the achievement gap between Chicago and other districts appeared to narrow. However, this trend began, at least in grades 3 and 6, before the introduction of high-stakes testing in these grades, and there was no noticeable break in the trend in 1997, the first year of the accountability system. Achievement scores in grade 8, particularly in reading, showed some break beginning in 1996 (the accountability policy began for 8th graders in 1996).

What can be inferred from these trends? On the one hand, a simple comparison of student achievement at the beginning and end of the decade suggests that Chicago experienced roughly comparable improvement on the IGAP and the ITBS. This might lead to the conclusion that the achievement gains on both exams were largely generalizable. On the other hand, a comparison of how achievement in the late 1990s changed in relation to the preexisting trends on each exam suggests that the accountability policy had a large effect on ITBS scores but little if any effect on IGAP scores. This might lead to the conclusion that the ITBS gains in Chicago were driven largely by test-score inflation.

The data do not necessarily support either conclusion, however. One prob-



Neither changes in student composition nor preexisting achievement trends in Chicago explain the substantial improvement in student performance since 1997.

lem is that the ITBS and IGAP are different in both content and format. In mathematics, the ITBS places more emphasis on computation, while the IGAP appears to give greater weight to problem-solving skills. Indeed, the computation items on the IGAP are often asked in the context of a word problem. The general format of the reading comprehension sections on the two exams are similar—both ask students to read passages and then answer questions about the passage—but the IGAP consists of fewer, but longer passages, whereas the ITBS contains a greater number of passages, each of which is shorter. Perhaps more important, the questions on the IGAP may have multiple correct responses in comparison with the ITBS, on which there is only one correct response. The fact that the two exams displayed different trends before the introduction of the accountability policy suggests that they

measure somewhat different concepts.

The natural solution would be to adjust the ITBS and IGAP scores to account for such differences in content. For example, one might estimate what the IGAP scores would have been if both exams had the same distribution of question types. In practice, this exercise is probably only feasible in mathematics, where test items can be categorized with relative precision. Moreover, this requires detailed item-level information for both exams, which is not available for the IGAP.

A second difficulty in interpreting the differences in performance trends between the two exams involves student effort. Students undoubtedly began to increase test-day effort for the ITBS after 1996. It is unclear how, if at all, effort levels changed on the IGAP. One might imagine that effort increased somewhat given the new climate surrounding testing. Equally plausible, however, is the idea that effort has declined now that teachers and students view IGAP scores as largely irrelevant. If student effort on the ITBS increased at the same time that effort on the IGAP decreased, one would expect more rapid achievement growth on the ITBS even if the exams were identical or learning were completely generalizable.

In sum, given the differences in composition between the two exams along with possible changes in student effort over the time period, it is extremely difficult to determine what one should expect to see under the best of circumstances.

Meaningful Improvement

It is important not to exaggerate the importance of the fact that gains may not generalize to other exams. Even if an accountability program produces true, meaningful gains, we would not expect gains on one test to be completely reflected in data from other tests because of the inherent differences across exams. Even the most comprehensive achieve-

ment exam can cover only a fraction of the possible skills and topics within a particular domain. For this reason, different exams often lead to different inferences about student mastery, regardless of whether any type of accountability policy is in place.

Yet in discussing how to interpret test-score gains, even testing experts occasionally slip into language that seems to neglect the value of gains in particular areas. Harvard scholar Daniel Koretz notes, “When scores increase, students clearly have improved the mastery of the sample included in the test. This is of no interest, however, unless the improvement justifies the inference that students have attained greater mastery of the domain the test is intended to represent.” Does this mean that if children improve their ability to add fractions, interpret line graphs, or identify the main idea of a written passage, this is of *no* interest?

Most people would agree that these improvements, while limited to specific skills or topics, are indeed important. This suggests an alternative criterion by which to judge changes in student performance—namely, that achievement gains on test items that measure particular skills or understandings may be *meaningful* even if the student’s overall test score does not fully generalize to other exams. To be meaningful, achievement gains must result from greater student understanding, and they must be important in some educational sense.

Test-score gains that result from cheating on the part of students or teachers would of course not be considered meaningful. Similarly, most people would not view as meaningful increases in performance that result from an improvement in testing conditions. A less clear-cut case involves student effort. Various studies have shown that accountability policies lead students to take standardized exams more seriously, either by working harder during the school year or by simply concentrating harder during the actual exam (or both). While the former clearly

represents meaningful gains, the latter may not. One could argue that teaching students to try hard in critical situations is a useful thing. But the observed improvements in student performance would represent greater effort rather than greater understanding.

A Close Look at the Questions

One way to assess the meaningfulness of reported achievement gains is to see how changes in student performance varied across individual test questions. While item analysis is not a new technique, it may provide important insight in assessing the effects of testing policies.

Consider test completion rates on the ITBS. Since there is no penalty for guessing on the ITBS (total score is determined solely by the number correct), the simplest way for a student to increase his or her expected score is to make sure that no items are left blank. Before the introduction of the accountability policy in Chicago, a surprisingly high proportion of students left one or more items of the ITBS exam blank. In 1994 only 58 and 77 percent of 8th grade

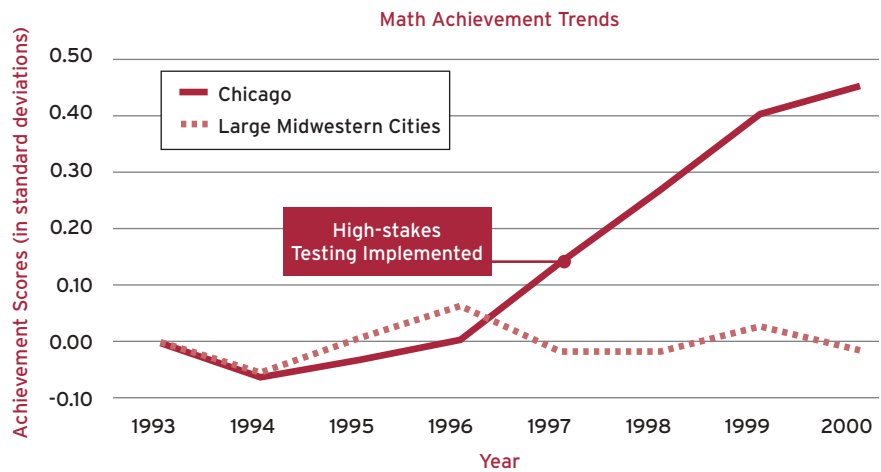
students completed the entire math and reading exams, respectively.

Test-completion rates increased sharply under the high-stakes testing regime. For instance, the number of 8th graders who completed the entire math exam increased to nearly 63 percent in 1998, with the vast majority of students leaving only one or two items blank. The greatest impact was for low-achieving students, largely because the overwhelming majority of higher-achieving students had completed the exam before the onset of high-stakes testing.

Can guessing explain the observed achievement gains in Chicago? If the increased test scores were due solely to guessing, the percent of questions answered would increase, but the percent of questions answered *correctly* (as a percent of all *answered* questions) would remain constant or perhaps even decline. In Chicago, the percent of questions answered has increased, but the percent answered correctly has also gone up, suggesting that the higher completion rates were not entirely due to guessing. A more detailed analysis suggests that guessing could explain only a small

Best in Show (Figure 2)

Chicago experienced stronger gains than other large urban school districts in the Midwest in the years after high-stakes testing was implemented.



Note: The achievement series for large midwestern cities includes data from Cincinnati, Gary, Indianapolis, St. Louis, and Milwaukee. The sample includes all grades from 3 to 8 for which data on test scores were available. Trends in reading were similar to mathematics.

SOURCE: Author

fraction of the overall achievement gains.

Next consider student performance across skill areas. The ITBS math section measures students' understanding of five broad areas: number concepts, estimation, problem-solving, data interpretation, and computation. Questions in the reading section are broken into three broad categories: understanding factual information, evaluating written material (identifying the author's viewpoint, determining the main idea), and drawing inferences (inferring the feelings, motives and traits of characters in a story, predicting likely outcomes).

The size of achievement gains following the introduction of the accountability policy differed across item areas. Students improved 7.1 percentage points on items involving number concepts and 6.8 percentage points on items involving computation. By contrast, students gained only 4.3 percentage points on problem-solving items and roughly 5.5 percentage points on data interpretation and estimation questions. Overall, these results suggest that math teachers may have focused on specific content areas in response to the accountability policy. Given the considerable weight placed on mathematical computation and number concepts in the ITBS, perhaps along with the perceived ease of teaching these skills, it would not be surprising if teachers chose to focus their energy in these areas. In reading, students made comparable improvement (roughly 5 percentage points) across question type, suggesting that test preparation may have played a larger role in math than in reading.

These aggregate results provide some insight, but the accountability policy affected students and schools differently based on previous achievement levels. In particular, observers have expressed concerns that the lowest-performing schools have responded to the policy by simply focusing on test preparation. If this were true, one would expect the patterns of test-score gains across items to differ for low- versus high-performing students and schools.

Low- and moderate-achieving schools made gains greater than those of high-achieving schools under the accountability policy.

To explore this, I examined achievement changes by item type for low-, moderate-, and high-performing schools, as measured by the percentage of students scoring at or above national norms on the ITBS reading exam in 1995. Schools with fewer than 20 percent of students meeting norms were defined as low achieving; those with 20 to 30 percent meeting norms were moderate achieving; and those with at least 30 percent of students' meeting norms were high achieving (this created three groups of equal size). Two patterns stood out. In both reading and mathematics, low- and moderate-achieving schools made overall gains greater than those of high-achieving schools under the accountability policy. This is consistent with the incentives generated by the policy, which placed low-achieving schools on probation. Regardless of previous achievement level, however, all schools appeared to have improved more in computation and number concepts than in other math concepts. Interestingly, while low-achieving schools improved in areas such as problem-solving and data analysis, higher-achieving schools made little if any improvement in these areas. For reading, regardless of school performance level, students showed similar improvement on items measuring factual, inferential, and evaluative understandings.

Conclusions

Chicago's experience with accountability

provides some lessons for other districts and states as they begin to implement the mandates of No Child Left Behind. The results of my analysis suggest that high-stakes testing substantially increases math and reading performance, with gains on the order of 0.20 to 0.30 standard deviations. Item-level analysis of test-score gains in Chicago during the 1990s reveals that math gains were disproportionately focused in certain areas, and therefore may not generalize to alternative performance measures, particularly those that tap other domains of knowledge. Nonetheless, this does not mean that the gains were not meaningful. They may well reflect an authentic increase in certain areas of knowledge and skills, underscoring the need for careful attention to the specific content of the exams used to hold schools and students accountable. Furthermore, it is important to note that the performance of Chicago's students on alternative assessments continued to increase in absolute terms, which may mean that there was no substantial tradeoff in skills learned.

The broader lesson is that educators and policymakers must look beyond aggregate measures of student performance to assess the nature of observed performance trends, and they must carefully distinguish between concerns of generalizability and concerns of meaningfulness. Overall, these results suggest that high-stakes testing has the potential to improve student learning meaningfully, but attempts to generalize the results to other learning must be approached with caution.

—Brian Jacob is an assistant professor of public policy at Harvard University's John F. Kennedy School of Government and an associate with the National Bureau of Economic Research (NBER) in Cambridge, Mass. This article is based on a paper presented at a June 2002 conference of Harvard's Program on Education Policy and Governance and an NBER Working Paper, "Accountability, Incentives, and Behavior."