

TABLE 2

Gain in Percent of Students at Basic Skills Level or Better, NAEP 8th-Grade Math, 1996–2000, as Function of 1996 Level and Accountability, Across States, by Race/Ethnicity

Independent variable	White Gain		Black Gain ^a		Hispanic Gain	
	I	II	I	II	I	II
Accountability index	1.134** (2.86)	1.41* (2.48)	1.77** (3.01)	2.57** (3.70)	3.17** (3.16)	4.47** (3.36)
1996 8th-grade math	-0.088 (1.11)	-0.140 (1.17)	0.211 (1.53)	0.052 (0.25)	-0.017 (0.12)	-0.142 (0.71)
Percent Black and Hispanic		-0.046 (0.01)		-14.53 (1.52)		-31.44 (1.96)
Population—July 1995		-0.00014 (-1.06)		-0.00017 (1.21)		0.00068 (0.25)
Average per-pupil revenue 1990		0.00083 (1.25)		0.00091 (1.04)		0.00102 (0.81)
Yearly growth in percent Black or Hispanic		2.39 (0.07)		-64.37 (1.16)		-54.49 (0.74)
Yearly population growth		-75.09 (-0.75)		161.19 (1.02)		15.66 (0.06)
Constant term	8.37 (1.43)	8.97 (1.27)	-4.20 (1.06)	-1.71 (0.32)	-0.685 (0.10)	4.54 (0.43)
R ²	0.25	0.34	0.3626	0.5546	0.27	0.41
Sample size	37	34	25	25	33	30

Note. With the alternative index, the coefficient and *t*-statistic on the index in column II for Whites, Blacks, and Hispanics are 1.26 (2.45), 2.43 (3.52) and 4.84 (4.25) respectively.

^a We omitted Nebraska from the models for Blacks because the Black scores were so low in 2000 and the number of test takers was low. With Nebraska included the coefficients on accountability for Blacks increase to 2.41 (3.88), 3.10 (4.29). *t*-values in parentheses. *p* < .10, **p* < .05, ***p* < .01, ****p* < .001.

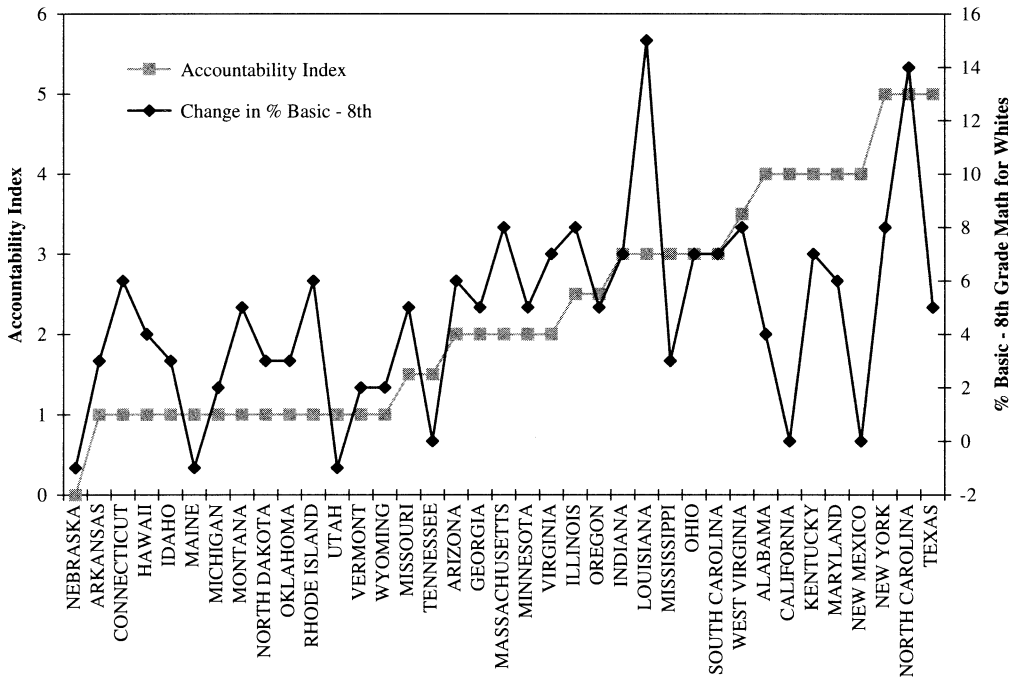


FIGURE 1A. Accountability index and the gain in the percent of 8th graders reaching the basic level on the NAEP math exam from 1996–2000, Whites.

TABLE 3

Gain in Percent of Students at Basic Skills Level or Better, NAEP 4th-Grade Math, 1996–2000, as Function of 1996 Level and Accountability, Across States, by Race/Ethnicity

Independent variable	White Gain		Black Gain ^a		Hispanic Gain	
	I	II	I	II	I	II
Accountability index	0.766 ⁻ (1.68)	.194 (0.29)	1.80* (2.35)	2.54** (2.94)	1.91** (2.95)	1.70 ⁻ (1.70)
1996 8th-grade math	-0.159 (1.55)	-0.268* (2.05)	-0.091 (0.59)	-1.59 (0.95)	-0.270** (2.92)	-0.337* (2.39)
Percent Black and Hispanic		6.03 (0.94)		-2.32 (0.24)		-10.70 (0.81)
Population—July 1995		0.000050 (0.36)		-0.00016 (0.97)		0.00023 (1.10)
Average per-pupil revenue 1990		0.0010 (1.48)		.0011 (1.11)		0.00043 (0.47)
Yearly growth in percent Black or Hispanic		-8.21 (0.22)		61.23 (1.09)		-46.39 (0.84)
Yearly population growth		-78.19 (0.69)		434.45* (2.38)		25.19 (0.15)
Constant term	14.81 (1.97)	18.28 (2.13)	6.53 (1.29)	-1.22 (0.20)	12.58 (2.79)	19.91 (2.16)
R ²	0.14	0.28	0.20	0.48	0.41	0.43
Sample size	36	33	25	25	32	309

Note. With the alternative index, the coefficient and *t*-statistic on the index in Column II for Whites, Blacks, and Hispanics are .403 (0.67), 2.43 (3.52), and 2.01 (2.18) respectively.

^aWe omitted Nebraska from the models for Blacks because the Black scores were so low in 2000 and the number of test takers was low. With Nebraska included the coefficients on accountability for Blacks increase to 2.51 (3.16), 3.18 (3.15). *t*-values in parentheses. ⁻*p* < .10, **p* < .05, ***p* < .01, ****p* < .001.

achieving at least the basic level. The coefficient of the accountability index is not significantly different from zero in the estimated equation for White 4th graders. Since Blacks and Hispanics start out at lower levels of basic skills proficiency than Whites, it may be easier to raise their low basic skills in the primary grades. This is partially borne out by our estimates. The estimated increase in the proportion of Hispanic 4th graders scoring at the basic skill level or higher corresponding to accountability strength is only marginally significant. The point estimate in the fully specified model suggests that a two-step increase in accountability would increase the percent achieving basic by 3.4 percentage points (just over half of a standard deviation in the change score). For Black students the impact of accountability is significant and suggests a 5.1% increase in basic skills with a two-step increase in accountability.

The effect of strong accountability systems at higher skill proficiency levels on the NAEP test might be expected to be less, given the relatively

“basic” nature of the test that most states use for accountability. We do, however, find a significant relationship between proficiency and the strength of the accountability system for all racial ethnic groups (Table 4). A two-step increase in the accountability index implies a 2.4 percentage point gain in the percent of White students and Black students scoring at proficient or better on the test, and 3.8 percentage point gain in the percent of Hispanic students scoring at proficient or better on the test.

Adjusting Results for Differential Exclusion and Inclusion Rates

A potentially serious bias in the NAEP math gains may arise because some students are eligible for exclusion from the test because they are designated as special education (SD) or limited English proficient (LEP). The proportion of SD plus LEP varies greatly among states. All states have some of these students take the standard NAEP test without accommodation and exclude others. A potential bias in gains arises because the

**Figure 1. State Accountability over Time
(with NAEP Testing Dates)**

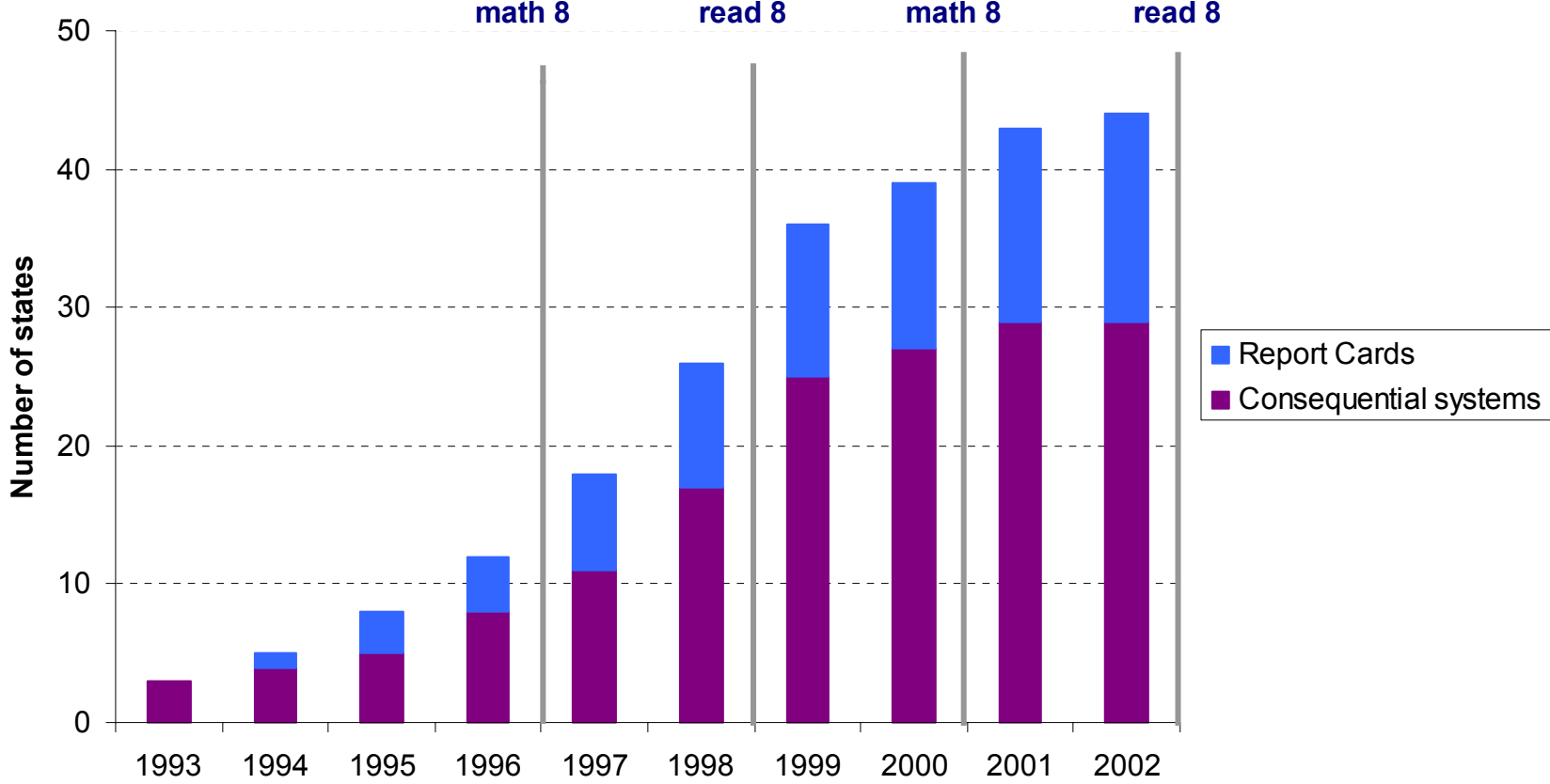
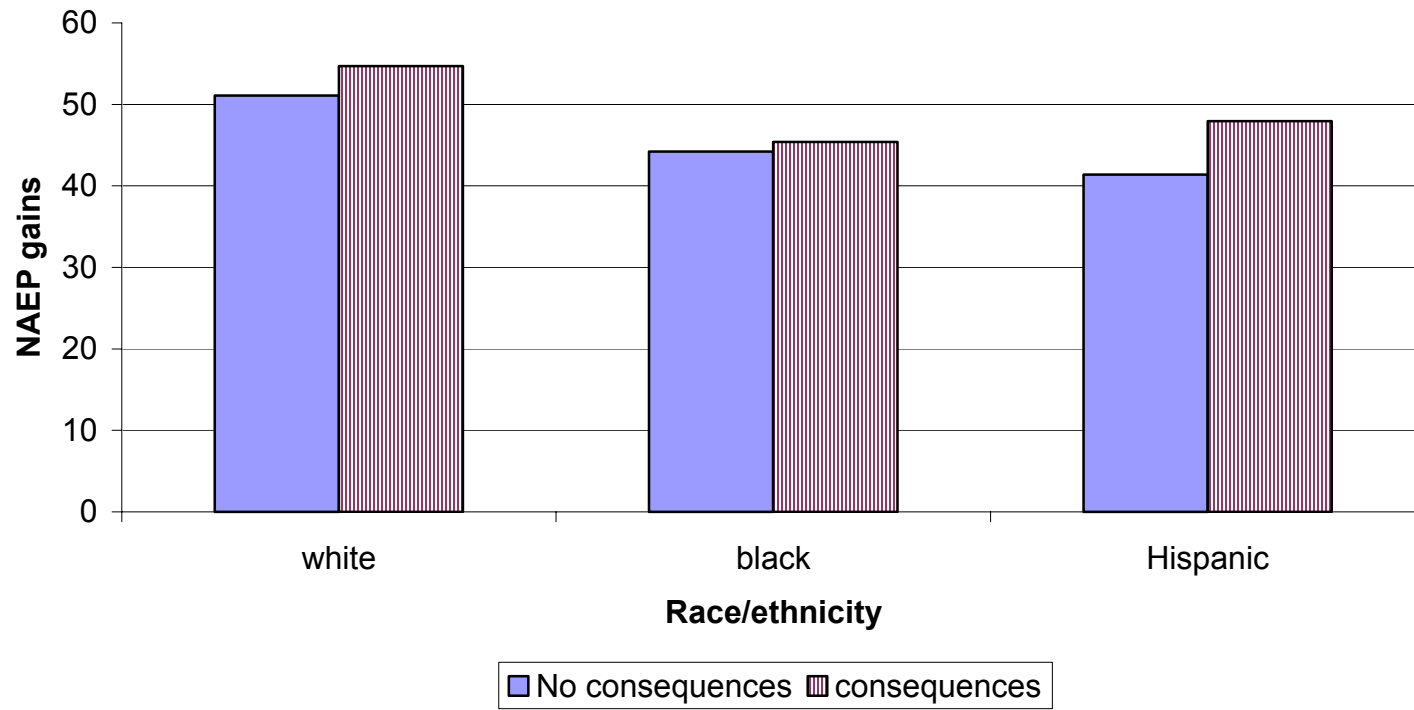


Figure 2. Effect of Consequential Accountability on Achievement by Race/ethnicity

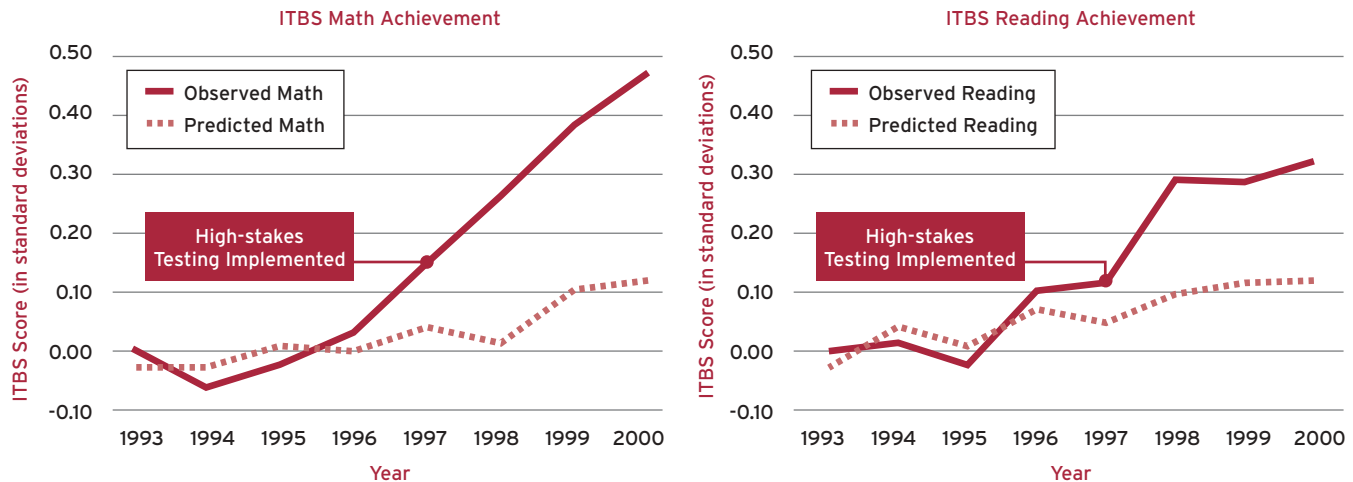


**Figure 3. Racial/Ethnic Gaps
by Consequential Accountability Status
(NAEP gains relative to whites)**



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

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

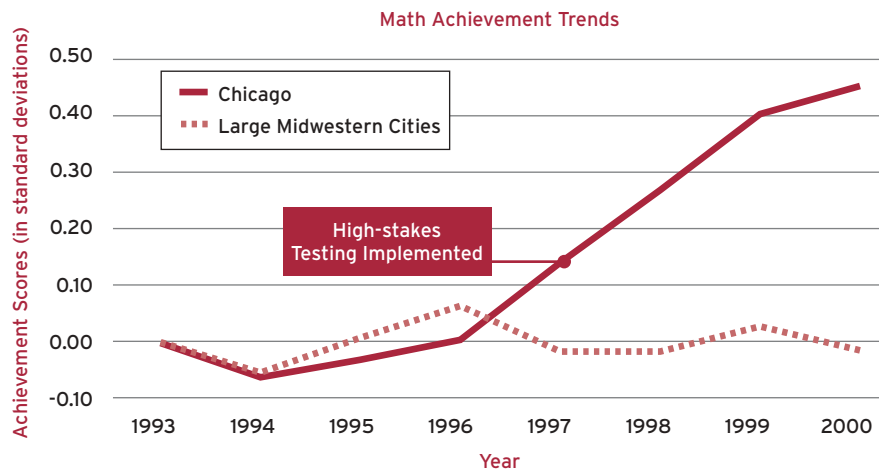
students completed the entire math and reading exams, respectively.

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