The Effects of High-Stakes Testing On Achievement: Preliminary Findings About Generalization Across Tests

Daniel M. Koretz
The RAND Corporation

Robert L. Linn
University of Colorado

Stephen B. Dunbar
University of Iowa

Lorrie A. Shepard
University of Colorado

Originally presented in R. L. Linn (Chair), *Effects of High-Stakes Educational Testing on Instruction and Achievement,* symposium presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April 5, 1991. This paper has been edited for clarity and differs in minor ways from the version originally presented.

The research reported here was supported by a grant from The Spencer Foundation and by the Center For Research on Evaluation, Standards, and Student Testing, which is partly supported by the Office of Educational Research and Improvement, U.S. Department of Education (OERI/ED). The opinions expressed here are solely those o the authors; no endorsement by The Spencer Foundation or OERI/ED is implied.

The authors wish to acknowledge the assistance of Ellen Harrison and Elizabeth Lewis with the data analysis report here.

Achievement testing has been a primary instrument of educational reform in the last decade. In many states and districts across the country, testing programs have been transformed from a means of monitoring the progress of students into a mechanism for holding students, teachers, and school administrators accountable. This test-based accountability, according to its proponents, will cause real improvements in the performance of the educational system and in the achievement of students.

Most observers agree that these changes have had profound effects, but a vehement debate has arisen about their desirability. Increases in test scores have bee observed in many states and districts during the first few years following the introduction of a high-stakes testing program. The degree to which those increases in scores reflect real improvements in student achievement, however--rather than gains specific to the particular test--has been the subject of intense debate. In addition, there has been a serious debate about whether the instructional changes caused by the test-based accountability are desirable or pernicious.

Based on some our previous research and observations (e.g., Koretz, 1988; Linn, Graue, & Sanders, 1990; Shepard, 1990) we expected that the rosy picture painted by results on high-stakes tests would be to a substantial degree illusory and misleading. That is, we expected that good performance on high-stakes tests would be caused in part by focusing undue attention on the specific content of the tests and therefore would not generalize very well when alternative measures of the same content and skills were used.

Evidence relevant to this debate has been limited. Cannell's reports on the "Lake Wobegon" effect (e.g., Cannell, 1988), in which he maintained that almost all states and most districts reported themselves to be "above average," began to give public credence to the view that scores on high-stakes tests could be inflated. Our follow--up investigation (Linn, Graue, & Sanders, 1990) suggested that the gains that were observed during the late 1980s were partially spurious and not supported by independent results from the National Assessment of Educational Progress. The study we are reporting today takes the next step, in that it provides detailed evidence from specific districts about the extent of generalization from high-stakes to other the and about the instructional effects of high-stakes testing.

Most of the comparisons I will present, which represent only our initial findings and are limited to grade three in one of our sites, suggest that performance does not generalize well from the district's high-stakes test to other tests that we administered. The disparities

among tests varied substantially from case to case but were generally considerable. The implication appears clear: students in this district are prepared for the high-stakes testing in ways that boost scores on that specific test substantially more than actual achievement in the domains that the tests are intended

measure. Public reporting of these scores therefore creates an illusion of successful accountability and educational performance. (Evidence pertaining to the second major issue, the effects of high-stakes testing on instruction, are presented in another paper in this symposium [Shepard, 1991].)

CHARACTERISTICS OF THE DISTRICT

I cannot describe the district or the tests that it has used in detail because we have agreed to protect the district's confidentiality.[1] However, some basic information is needed in order to interpret the results that I will present. The district involved--"District B" in our larger design--is a large, high-poverty urban district with large numbers of both black and Hispanic students. Three-fourths of our sample schools had non-Asian minority enrollments of 70 percent or more, and about half minority enrollments above 90 percent. Three-fourths of the third-grade students in our median school received free lunch. The district as a whole was very similar to our sample in these respects.

The district's overall minority enrollment did not change much during the four years of our study, but the ethnic composition of some schools changed markedly. The minority enrollment in the typical school in our sample changed by less than 2 percentage points between 1986 and 1990. The range, however, was from a 20-percentage point decrease to a 23-percentage point increase.

The district uses unmodified commercial achievement tests for its testing program, which is perceived as high-stakes. Through the spring of 1986, they used a test that I will call Test C. Since then, they have used another, called Test B, which was normed 7 years later than Test C. The district publishes school median grade equivalents (GEs) on the tests. Our median schools had median scores on Test B in 1990 that were about average in vocabulary (GE = 3.6), below average in reading (GE = 3. 1), and above average in mathematics (GE = 4.3; see Table 1). (The district tests in spring, so a GE of 3.7 corresponds to the 50th percentile.) Because the distribution of school medians on the district's test is positively skewed, the means of school medians were 1 to 5 academic months higher than the medians.

---

[1] Although we cannot credit the individuals who cooperated in this study by name, we gratefully acknowledge their assistance. A study of this sort is both burdensome and politically risky, and a good many districts decided that they could not participate for those reasons. Indeed, one of our primary sites pulled out of the study weeks before we would have administered tests, citing the political risks they would face if the district were identified despite our efforts to keep its identity confidential. The individuals in our participating sites deserve credit for shouldering the risk and burden that this study involved.

Table 1. School Medians on District's High-Stakes Test (Test B), GEs, 47 Schools

	Mean of School Medians	Median of School Medians
Math	4.4	4.3
Reading	3.6	3.1
Vocabulary	3.9	3.6

DESIGN AND SAMPLE ADJUSTMENTS

We sampled intact classrooms from within schools on a random basis. Sampled schools were randomly divided into subsamples, and one subsample of schools was administered Test C, the test that the district had used through 1986. Other subsamples were administered a parallel form of Test B or alternate tests constructed to match Test B or Test C in content but not in format.

Because some of the district results to which we must compare our results are reported on the level of school buildings, it was necessary to check the degree to which our classrooms were typical of the schools from which they were drawn. We did this by comparing the school-wide results on the district's test to the performance on that same test of our within-school samples. This discrepancy varied markedly and was large in some schools. On average, the discrepancy was small, but students who were tested on our Test C tended to score better than the other students in their schools on the district's own Test B.

All school-level results based on Test C were adjusted accordingly. (Student-level results, which I will report later, did not require adjustment because each contrast is based on a single sample of students.) The adjustment required two steps. District B reports its results in median grade equivalents, but the Test B and Test C GE scales are quite different metrics. The Test B GE scale has a great deal of positive skew. As a result, adjusting our Test C GEs for the Test B GE difference between our samples and their schools could have biased our results. (In practice, doing so would have favored our hypotheses.) Instead, we calculated the differences on Test B in terms of national percentile ranks. We then mapped those onto the Test C GE distribution using the publisher's norms and adjusted our Test C medians for the difference.

HISTORICAL TRENDS

For this analysis, we compared the district's own results--for Test C in 1986 and for Test B in 1987 through 1990--to our results for Test C. Our Test C results reflect 840 students in 36 schools.

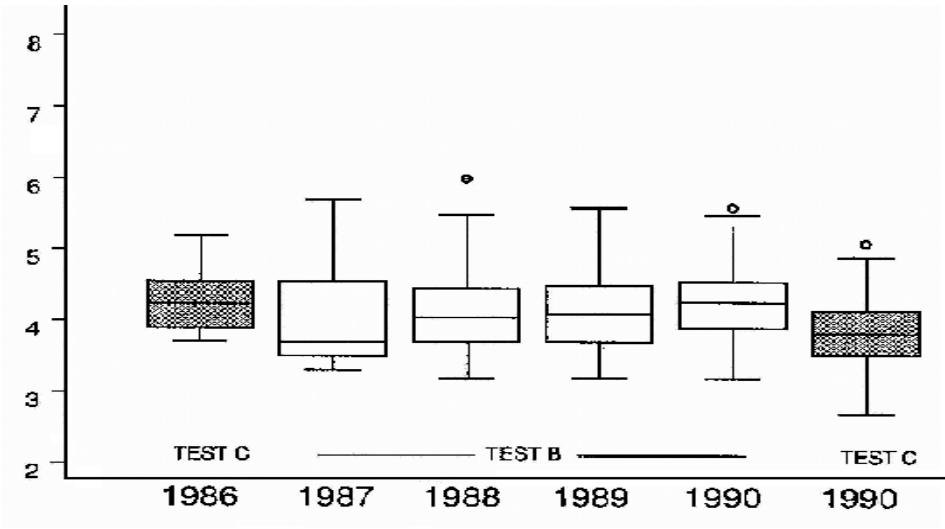
Mathematics

The results in mathematics show that scores do not generalize well from the district's test to Test C, even though Test C was the district's own test only four years ago and is reasonably similar in format to Test B. (That is, both Test C and Test B are conventional, off-the-shelf multiple-choice tests.) The school-level results on which this conclusion rests are displayed for mathematics (total) in Figure 1.

Looking at the first five medians starting with 1986, one sees the traditional pattern: scores dropped markedly when the new test was introduced and then rose again, particularly between the first and second years of the new test. The drop when the test was changed was about half an academic year: from a GE of 4.25 to 3.7 This drop presumably reflects two factors: lesser familiarity with the test and more recent, and hence harder, norms.[2]

[2] The norms became more difficult on both Test A and Test B, although not by equal amounts.

Figure 1. Distribution of Median Math GE



Our re-administration of Test C showed that in mathematics, the schools in this sample had slipped by roughly four academic months in the four years since the district had itself used Test C. This can be seen by comparing the right-most box in Figure 1 to the left-most box. In 1990, the median school had an adjusted median of 3.83, compared to the median of 4.25 in the last year that Test C was the district's own test. For the spring of grade 3, a drop of this magnitude is not inconsequential.

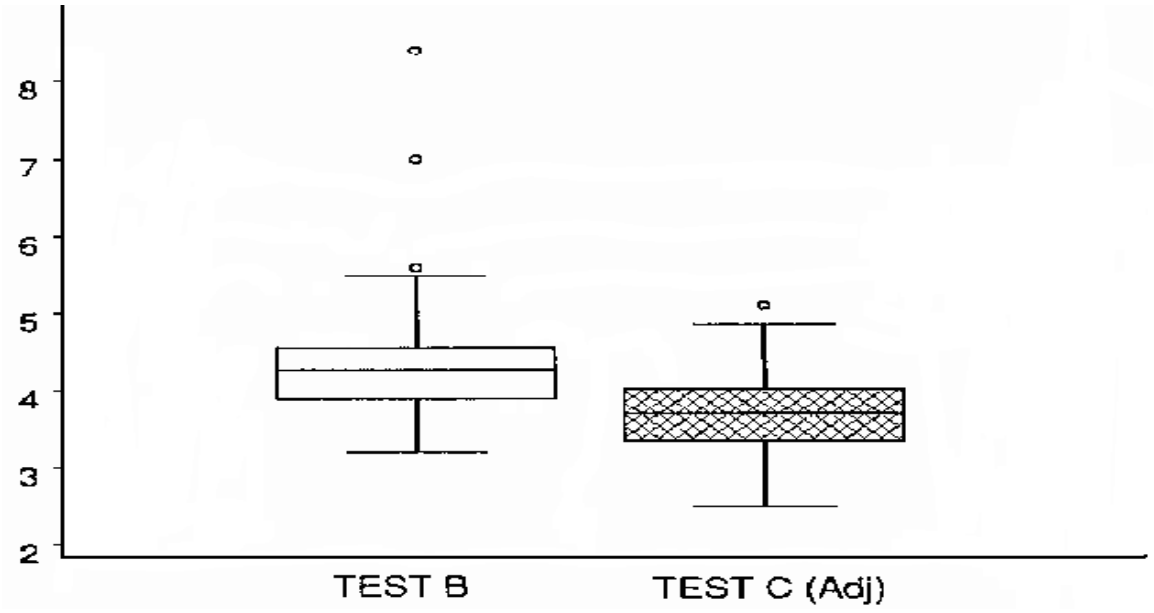
Our re-administration of Test C showed that in mathematics, the schools in this sample had slipped by roughly four academic months in the four years since the district had itself used Test C. This can be seen by comparing the right-most box in Figure 1 to the left-most box. In 1990, the median school had an adjusted median of 3.83, compared to the median of 4.25 in the last year that Test C was the district's own test. For the spring of grade 3, a drop of this magnitude is not inconsequential.

While the results we have just presented are appropriate for comparing scores on Test C in 1990 and 1986 (because they are based on the identical test and norms) they understate the discrepancy between our 1990 Test C results and the district's 1990 results on Test B--that is, the right-most two boxes in Figure 1. The reason is that the edition of Test C that the district used in 1986, and that we re-administered 1990, was normed 7 years earlier than the district's Test B. Accordingly, we adjusted our Test C medians using the publisher's conversion tables to newer norms for Test C that were set within one year of those used for Test B.

The effect of this adjustment is to further increase the discrepancy between our results and the district's own, although not dramatically. Schools' median scores in mathematics drop a month or two when placed on the newer Test C scale. With that adjustment, the median school had a median GE score of 3.7 on Test C, about 6 academic months lower than the median school on Test B (see Figure 2).

While these comparisons are bleak enough in terms of medians, they are more extreme yet if one considers the schools that score particularly well on the district's own Test B. While it would be easy to read too much into a distribution of only 36 school medians, the pattern is too striking to ignore. The distribution of medians on the district's test has more positive skew than does the distribution on Test C, and the top of the distribution on Test B is a GE of 8.4, more than three academic years above the highest median on our Test C (GE = 5. 1). To some degree, this may be function of the tests; the GE scale on Test B has a substantial positive skew, and the skewness of the school

Figure 2. Distribution of Median Math GE 1990 Cross-Sectional Comparison

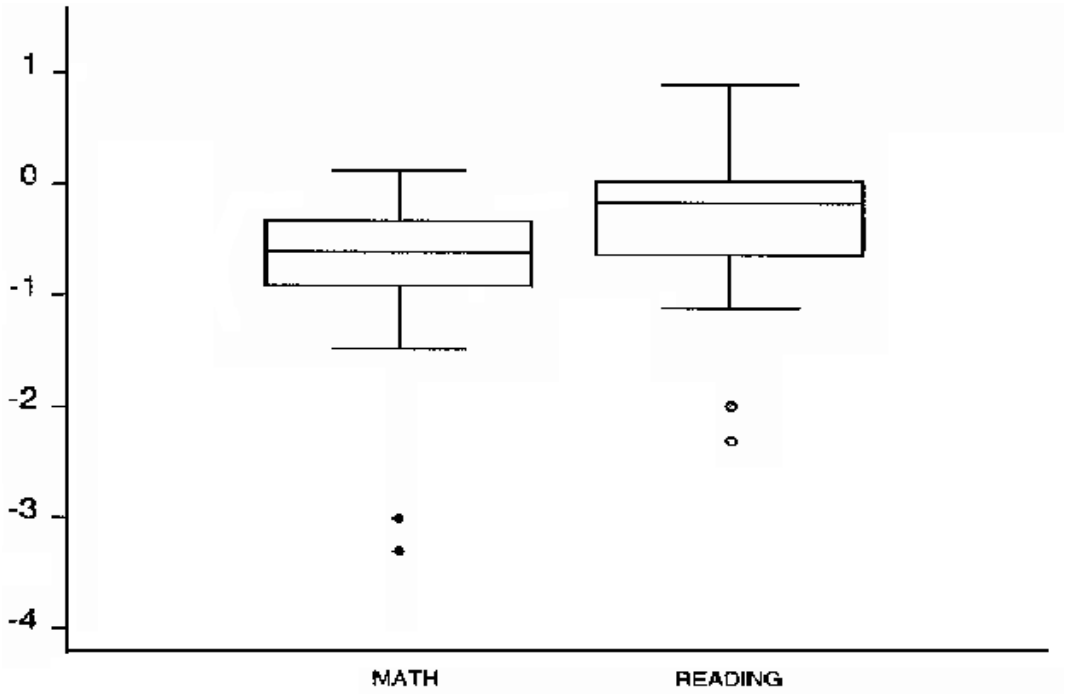


medians was apparent to some degree even in the first years the test. The movement of outliers to ever-higher values in subsequent years, however, suggests that test preparation may also be at work.

One effect of this skewness is that the *mean* of school medians shows an even greater difference between Tests B and C: nearly 8 academic months, compared to the 6-month difference between medians.

Another, more striking effect is that some schools showed enormous cross-sectional differences between Test B and our adjusted Test C. The distribution of individual school's cross-sectional differences in mathematics can be seen in Figure 3. The school that fared worst in this cross-sectional comparison scored a full three years and three months lower on Test C, and 5 of the 36 schools scored a year or more lower on Test C. By contrast, only 5 schools showed positive differences between Test C and Test B in 1990, and the largest of these differences was about 3 months.

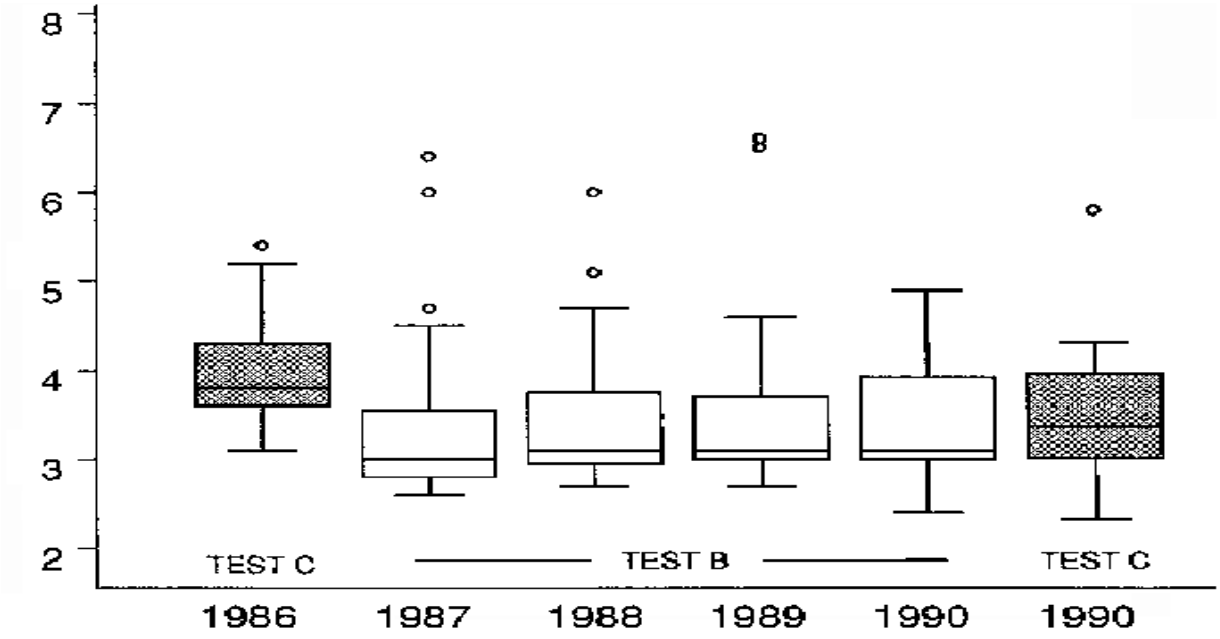
Figure 3. 1990 Cross-Sectional Differences Median GE: Test C - Test B



Reading

In terms of two of our contrasts, the lack of generalization was as bad or even worse in reading than in mathematics. The performance of our schools in reading slipped even more dramatically when Test B was first introduced. The median of school medians dropped by a full 8 academic months (Figure 4), from about average (GE = 3.8) to well below average (GE = 3.0). The comparison of our Test C the district's final administration of that test in 1986 shows essentially the same that we found in mathematics: about 4 academic months, to a GE of about 3.4.

Figure 4. Distribution of Median Reading GE

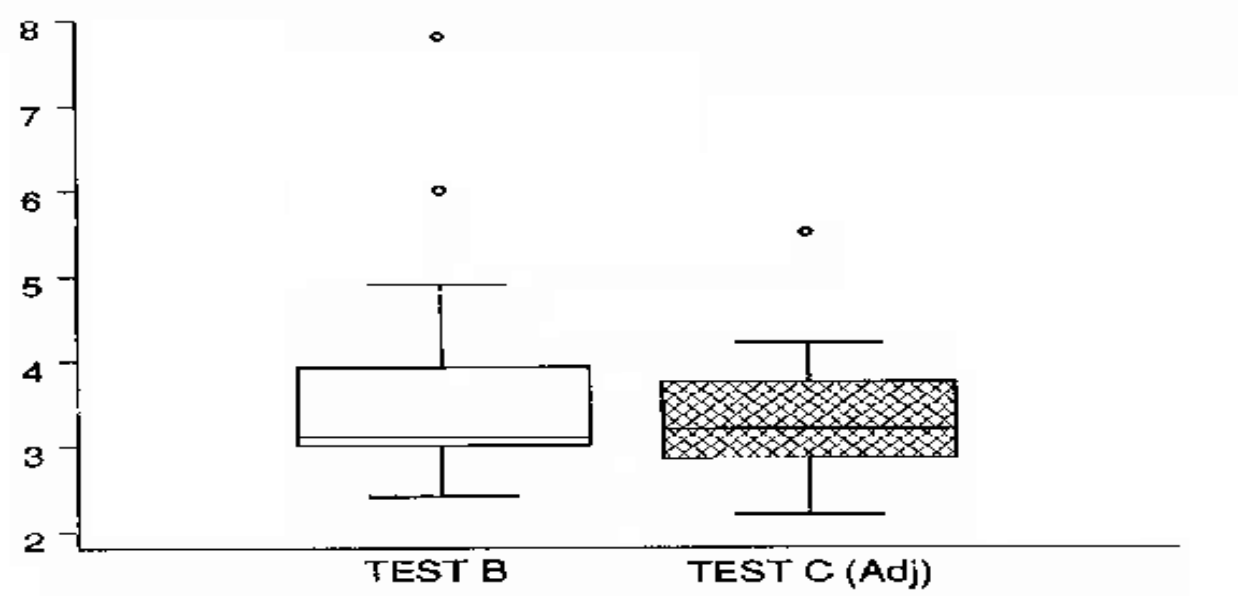


Our cross-sectional comparison of 1990 Test B and adjusted Test C results, however, showed relatively little difference in terms of the median schools. The *median* of school medians was in fact *higher,* albeit trivially, on our adjusted Test (by less than one academic month; see Figure 5). Here again, however, the distribution of schools was more positively skewed on the district's Test B, and accordingly the *mean* of school medians showed a moderate drop--roughly three academic months, from Tests B to Test C.

Thus we have an inconsistency: in terms of change over time, lack of generalization across tests is at least as great in reading as in mathematics, while in terms of cross-sectional comparisons, the disparity

is in reading is smaller or nonexistent, depending on the measure. This inconsistency hinges in part on the differences in trends on the district's own Test B. In mathematics, as noted earlier, the median school's score followed the trajectory we would expect: a sharp drop when the district changed tests, followed by a fairly rapid recovery as everyone gets used the new test. In reading, however, the recovery never happened. The median school score on the district's Test B rose only one month between 1987 and 1990, and the of school medians rose only 2 months.

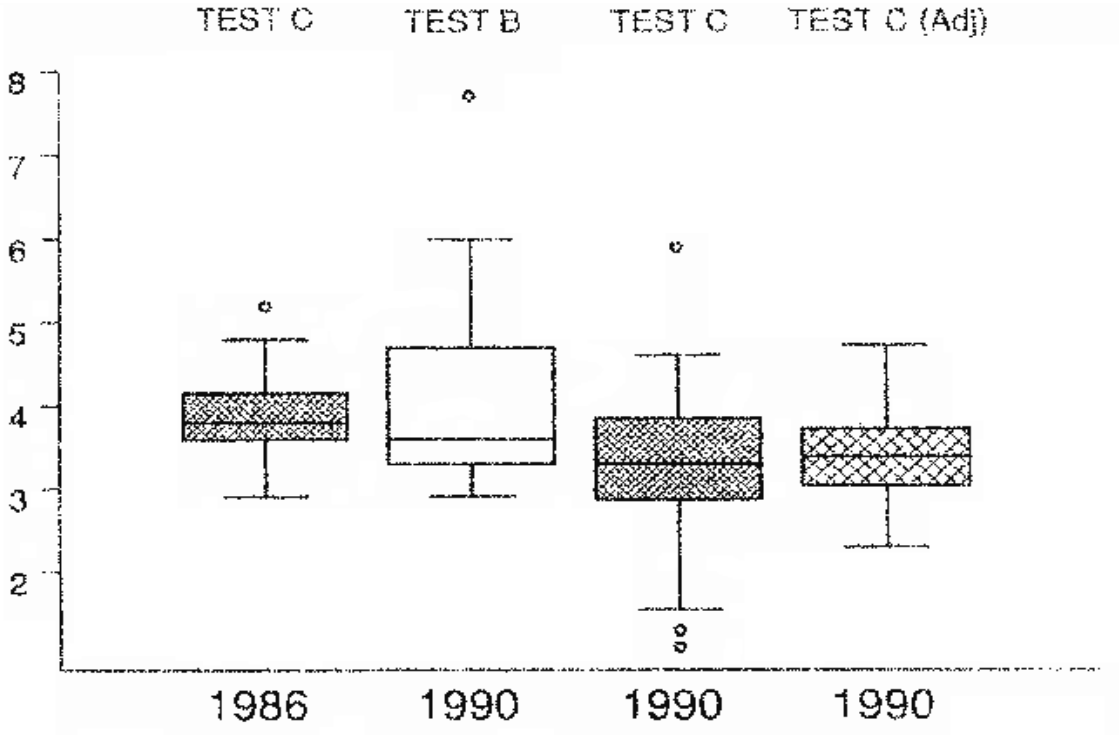
Figure 5. Distribution of Median Reading GE 1990 Cross-Sectional Comparison



Vocabulary

Our sample showed a modest weakness of generalization across tests in vocabulary as well. The median school was about average on Test C when the district last used it in 1986, with a median GE of 3.8. In 1990, the median school scored two months lower on the district's Test B (GE = 3.6: see Figure 6). On our re-administration of Test C, however, the median school scored nearly four months lower than the median on that test 4 years earlier (GE = 3.4). When Test C is scored against the more recent norms, the cross-sectional difference between the median schools for Tests B and C remains about two months.

Figure 6. Distribution of Median Vocabulary GE



RELATIONSHIPS BETWEEN CHANGE AND OTHER FACTORS

We have conducted limited analyses of the relationships between historical change on Test C and other variables. These analyses are consistent with our primary hypothesis of inflation of scores from teaching to the test but are inconsistent with on of our secondary hypotheses.

One of our secondary hypotheses was that scores would tend to be more inflated in the case of lower-scoring minority students. In later work, we will be able to address that question more directly, because we will be building longitudinal records at the student level. As a first look at that question, however, we examined the relationship between the ethnic composition of our schools and the change in median Test C scores. We found that the relationships between change on Test C and both overall percent minority and percent receiving free lunch were all very small, essentially zero in most cases.

Any relationship between the drop in Test C and *change* in minority enrollments or poverty is important for another reason: the possibility that scores dropped on Test C in some schools not because of teaching to the test, but rather because of demographic change. The demographic change that we can measure may in fact account for some of the variation among schools in the decline on Test C. Pearson

correlations between changes on Test C and change in percent minority enrollment ranged from -.24 (for vocabulary) to -.31 (for mathematics). This relationship cannot account for the overall decline on Test C, however, because on average, our schools showed almost no change in minority enrollments. Change on Test C was not appreciably related to change in percent receiving free lunch.[3]

STUDENT-LEVEL COMPARISONS

We have only begun the tests of generalization of performance across tests at the level of individual students, but at this point we have comparisons involving three of our tests to the district's Test B. All of these comparisons are based on the performance of individual students on two tests. These contrasts are more straightforward than the school-level comparisons just reported because many confounding variables are eliminated.

The first student-level comparison involved administering a parallel form of district's Test B to a sample of students a few weeks after the district's own testing. This was included in the design more for methodological than for substantive reasons, but I report the results here because they remove certain threats to the validity of our conclusions.

Three factors enter into any differences between the district's test and parallel form:

- motivational differences (a lack of interest in performing well on our tests because they don't have consequences);
- practice effects (because we administered our parallel form later to avoid inappropriate practice for the district's test); and
- teaching to the specific items in the district's form of Test B.

The first and third of these would lead to lower scores on the parallel form, while the second would tend to produce higher scores. Possible bias from motivational factors were our primary concern and were the reason for incorporating a parallel form in our design.

Given that we cannot disentangle these three factors and that the first could lead us to make Type I errors, the ideal result from our perspective was to obtain reasonably similar results from the parallel form and the district's administration of Test B. That is what we found. All of our parallel-form comparisons were within a range of one academic month or two percentile points (Table 2). (Note that the medians here are higher than the school-level statistics reported in Table 1. This reflects the fact that our parallel-

[3] Data for free lunch receipt in 1990 have not been obtained yet, so the change in percent receiving free lunch is from 1986 to 1989.

test subsample, while randomly drawn, turned out not to be entirely representative of our sample as a whole.)

Table 2. Comparisons of Test B Scores and Parallel Form Scores, Median Percentile Ranks and GEs

	Reading (N = 133)		Mathematics (N = 136)		Vocabulary (N=136)	
	PR	GE	PR	GE	PR	GE
Test B	57	4.2	78	4.9	61	4.7
Parallel Form	55	4.1	78	4.8	63	4.8

The first of our substantive comparisons of student-level results contrasts score on one of our "alternate" tests to the district's Test B. we constructed two of these alternate tests. ("We" in this case is primarily the three people who follow me today: Lorrie Shepard, Roberta Flexer, and Elfrieda Hiebert.) You will hear more about these tests in the following papers. Each of the alternate tests was designed to parallel the conventional achievement test used in one of our districts in terms of content but not format. Thus these tests included item types such as open-ended questions and multiple-choice questions which permitted more than a single correct answer. One of these tests, which we call Alternate Test B, was designed to match District B's test and curriculum framework.

We assumed from the outset that to some degree, students would perform more poorly on our tests as a result of more difficult item types, regardless of any effects teaching to the test. Accordingly, our design called for equating each of our alternative tests in two other districts. The alternative test used in the district I am reporting today was equated to Test B using samples from one district where testing low stakes and a second district that has high-stakes testing but does not use Test B.

Comparison of students' performance on the equated alternate test and the district's Test B showed a substantial deficiency of generalization, particularly in mathematics. In mathematics, the students who took both of these tests--who do not overlap with the sample who took our re-administered Test C--scored 15 percentile points and 7 academic months lower on our alternative test than on the district's test (Table 3). In reading, the discrepancy was only about half as big but was nonetheless considerable: 7 percentile points and 3 academic months. (Note that the subsample administered the Alternate B was more similar than our parallel form subsample to sample as a whole.)

Table 3. Comparisons of Test B Scores and Equated Alternate Test B Scores, Mean Percentile Ranks and GEs

	Reading (N = 620)		Mathematics (N = 707)	
	PR	GE	PR	GE
Test B	42	3.4	35	3.1
Alternate Test	61	4.3	46	3.6

A second substantive comparison of student-level results contrasts the scores on Tests B and C for students who took both in 1990. Roughly 750 students in 34 schools are included in these results, depending on the scale.

In mathematics, the results of these student-level comparisons are quite similar to the cross-sectional school-level results already reported: they show a striking weakness of generalization. The median student in this subsample received a GE of 4.5 on the district's Test B (slightly higher than the median of our entire sample). This corresponds to a national percentile rank of 67 (Table 4). These same students' scores on Test C, once adjusted to the newer norms, were 7 academic months and 16 percentile points lower. (Note that the subsample administered Test C were similar to our total sample in reading but scored somewhat higher in mathematics.)

Table 4. Comparisons of Median Student Scores on Tests B and C, GEs and National Percentile Ranks

	Reading (N = 133)		Mathematics (N = 136)		Vocabulary (N=136)	
	PR	GE	PR	GE	PR	GE
Test B	42	3.4	67	4.5	48	3.6
Test C[a]	38	3.4	51	3.8	35	3.4

The results of this student-level comparison were quite different in reading: the median student scored as high on Test C as on the district's Test B in terms of GEs but slightly lower in the metric of national percentile ranks. (The estimated relationship between percentile ranks and GEs is not the same for Tests B and C). Recall that the school-level cross-sectional 1990 comparison between Tests B and C in reading also showed an atypically small discrepancy between the two tests.

[a] Test C results are expressed with reference to new norms that are within one year of those used for Test B.

The student-level comparison of vocabulary scores was consistent with the school-level results already reported. The median student scored two months lower on Test C than on the District's Test B (GEs of 3.6 and 3.4, respectively). The difference in terms of national percentile ranks, however, was more substantial: the median student scored at the 48th percentile on Test B but only the 35th percentile on Test C.

CONCLUSIONS

In mathematics, then, all of the comparisons presented here strongly support our primary hypothesis that performance on a conventional high-stakes test does not generalize well to other tests for which students have not been specifically prepared. Three of the five primary contrasts reported here showed differences in performance of six to eight academic months between the high-stakes test and others, the fourth was just shy of that, and the fifth showed a difference of four months. In terms of estimated percentile ranks, two of the contrasts showed differences of 15 or 16 percentile points.

The evidence in reading is less consistent but nonetheless suggests significant weaknesses of generalization in some instances. The historical comparison on Test C showed a fall-off of four academic months and the change from Test C to Test B in 1987 caused a drop of eight months. Our alternative test suggests a difference about half that large. The cross-sectional comparisons of Tests B and C are the exception both the school- and student levels: they show differences ranging from near zero to thee academic months, depending on the measure.

The more consistent and generally larger disparities among tests in mathematics are not surprising. Aggregate data on the "Lake Wobegon effect" show more inflation of scores in mathematics than in reading (Linn, Graue, and Sanders, 1990), and we therefore hypothesized that we would find weaker generalization in mathematics.

There is more to be done to explore this lack of generalization. Subsequent members of this panel will provide several other pieces of the puzzle: evidence about teachers' activities to prepare students for testing and item-level exploration of the disparities in performance between the district's Test B and our Alternate B. In coming months, we will be extending this work in several other ways: examining results from additional sites, exploring longitudinal patterns of change at the student level, and contrasting multiple-choice tests at the level of individual items and clusters of items.

Even the preliminary results we are presenting today, however, provide a very serious criticism of test-based accountability of the sort used in this site and in many other districts and states around the country. First, it suggests that the information provided to the public by accountability-oriented tests can be seriously misleading. Few citizens or policymakers, I suspect, are particularly interested in

performance, say, on "mathematics as tested by Test B but not Test C." They are presumably much more interested in performance in mathematics, rather broadly defined. Our preliminary results indicate that as a guide to performance in the domain in question, the results of this district's high-stakes test overstate achievement by as much as 8 academic months by the spring of grade 3.

Second, our results raise serious concerns about the effects of high-stakes testing on instruction. The past several years have seen continuing debates about appropriate and inappropriate teaching to the test. Skeptics about test-based accountability, including several of us, have suggested that undesirable narrowing of instruction is one likely consequence of high-stakes testing. Supporters of test-based accountability, on the other hand, argue that focusing on the content of the test is desirable, as long as test-based accountability leads teachers to focus on broad areas of knowledge and skills measured by the test rather than on content specific to the test question. Our results seem clear enough: to a substantial degree, teachers in this district must be focusing on content that is specific to the particular test used for accountability, rather than trying to improve achievement in the broader sense that we would all desire.

REFERENCES

Cannell, J. J. (1988). National normed elementary achievement testing in America's public schools: How all-50 states are above the national average. *Educational Measurement: Issues and Practice*, *7* (2), 5-9.

Koretz, D. (1988). Arriving at Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? *American Educator, 12*, (8)-15, 46-54,

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

Shepard, L. (1990). Inflated test score gains: is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice,* 9 (3), 15-22.

Shepard, L. (1991). The effects of high-stakes testing on instruction. In R. L. Linn (Chair), *Effects of High-Stakes Educational Testing on Instruction and Achievement,* symposium presented at the annual meeting of the American Educational Research Association and the National Council on Measurement Education, Chicago, April 5, 1991.