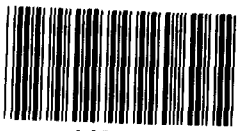
148850

GAO

United States
General Accounting Office
Washington, D.C. 20548

Human Resources Division

B-252634



March 30, 1993

The Honorable Dale E. Kildee
Chairman, Subcommittee on Elementary,
Secondary, and Vocational Education
Committee on Education and Labor
House of Representatives

Dear Mr. Chairman:

This letter presents information on the extent to which Chapter 1 schools may exit program improvement after 1 year on the basis of sufficient gains in students' achievement-test scores. This phenomenon is known as "testing out." Previous studies have reported that a significant proportion of schools identified for program improvement test out 1 year later--often without fully implementing an improvement plan.

We recently reported that imprecision associated with the achievement-test scores used to identify schools for program improvement can lead to inaccurate judgements about Chapter 1 program effectiveness.[1] Our conclusions were based, in part, on a statistical analysis of achievement-test data for Chapter 1 students in one large state. During a briefing of your staff, on January 22, 1993, concerning that report, we agreed to further analyze these data to determine (1) the percentage of schools that would have tested out after 1 year and (2) the extent to which these schools might be inaccurately judged as effective or ineffective because of imprecision in achievement-test scores.

We found that depending on the standard used for minimum achievement-test gains, one-half to three-quarters of the schools in our analysis that would have been identified in one year would also have tested out the following year. However, when we accounted for the imprecision associated with achievement-test scores, we found that a majority of these schools would have been initially identified on the

[1]Chapter 1 Accountability: Greater Focus on Program Goals Needed (GAO/HRD-93-69, Mar. 29, 1993).

056322/148850

basis of test scores that do not show conclusively that their students fell short of the standard for average achievement gains. In addition, a majority of these schools would have tested out with achievement-test scores that do show conclusively that their students exceeded the standard for achievement gains. Thus, for many of the schools in our analysis that would have tested out, we cannot be highly confident that they should have been identified the first year, but we can be highly confident that they exceeded the standard for achievement gains in the second year. (Our findings are presented in greater detail in enclosure I. Additional, supporting tables are presented in enclosure II.)

To conduct our analysis, we simulated the school identification process over 2 successive years, using a data set that contained achievement-test scores for Chapter 1 students in 1,684 schools in Pennsylvania. Because of certain limits on the scope of our analysis, however, our findings do not indicate the number of Chapter 1 schools in Pennsylvania that actually tested out during the time period we studied. (The scope and methodology of our review are described in enclosure III.)

We hope our review provides additional perspective on the process of identifying schools for Chapter 1 program improvement. Copies of this letter will be provided to the Secretary of Education and all state Chapter 1 coordinators; copies will also be made available to others on request. This review was conducted under the direction of Ruth Ann Heck, Assistant Director, who may be reached at (202) 512-7012 if you have any questions.

Sincerely yours,

Linda Morra

Linda Morra
Director, Education and
Employment Issues

Enclosures - 3

GAO'S ANALYSIS

We analyzed the extent to which schools in one state would have exited program improvement after 1 year on the basis of sufficient gains in their Chapter 1 students' average achievement-test scores. This phenomenon is known as "testing out." Before presenting our findings, however, we provide background information on how Chapter 1 program effectiveness is measured, the imprecision associated with this process, time frames for program improvement activities, and previous estimates of the extent of testing out.

BACKGROUND

Chapter 1 is the largest federal education program for children in elementary and secondary schools. The statutory goals of Chapter 1 are to help educationally deprived children[2] (1) succeed in the regular program of the school district, (2) attain grade-level proficiency, and (3) improve their achievement in basic and more advanced skills. To ensure that schools are effective in helping Chapter 1 students achieve these goals, the Congress created a new accountability system through the program improvement provisions of the Hawkins-Stafford Elementary and Secondary School Improvement Amendments of 1988 (P.L. 100-297).

Measuring Chapter 1 Program Effectiveness With Achievement Tests

Under the program improvement provisions, schools are evaluated on the average change in their Chapter 1 students' achievement-test scores over a 1-year period.[3] Students' scores are reported in terms of normal curve equivalents (NCEs), a special scale used for achievement testing in Chapter 1. The difference between students' average NCE score one year and the next is referred to as an "NCE change score." For example, if the Chapter 1 students in a school had an average NCE score of 32 NCEs one year and 34 NCEs the next year, then the school's NCE change score would be 2 NCEs. When a school's NCE change score is positive, its Chapter 1 students are said to have made gains in achievement; when it is negative, the students are said to have shown losses in achievement; and when it is 0, they are said to have maintained the same achievement level.

---

[2]An educationally deprived child is one whose educational attainment is below the level appropriate for his or her age.

[3]When we refer to "achievement tests," we mean standardized, norm-referenced, multiple-choice achievement tests.

To meet the federal minimum standard for achievement gains, schools must have an NCE change score greater than 0. We refer to this as the "0 NCE standard." However, states and districts are free to establish higher standards, such as requiring schools to make gains greater than 2 or 4 NCEs.

Achievement Tests and NCE Change Scores are Imprecise Measures

The achievement tests students take are not perfectly reliable. If a student took the same test several times, his or her score would not be the same every time. For example, a student might score 37 NCEs one time and 35 the next. On repeated testing, a student's scores would cluster around his or her hypothetical "true" score. Similarly, NCE change scores are imprecise measures of average changes in Chapter 1 student performance on achievement tests. NCE change scores can be affected by random fluctuations in students' scores and by the number of students included in the calculation. Thus, NCE change scores do not measure precisely the true gain or loss in achievement among a school's Chapter 1 students.

It is possible to account for this imprecision, however, by calculating the range within which a school's true NCE change score would fall and comparing this range to the NCE standard. This technique enables us to determine whether a school's NCE change score is above or below the NCE standard by a statistically significant margin. If the NCE standard falls outside this range, we say the school has a "conclusive" NCE change score because we can be highly confident that the school's true NCE change score is either above or below the standard. If the NCE standard falls within this range, we say the school has an "inconclusive" NCE change score because we cannot determine confidently whether the school's true NCE change score is above or below the standard.

Time Frames for Developing and Implementing Program Improvement Plans

After a Chapter 1 school is initially identified as in need of improvement, district and school officials may take up to 1 full school year to develop an improvement plan, although parts of this plan must be implemented as soon as possible. However, if the school is judged effective a year after it was initially identified, it may exit program improvement without completing or implementing its improvement plan. A previous study reported that when local officials believe a school has been inaccurately identified, they sometimes delay development and implementation of

an improvement plan in the hope that the school will test out the following year.[4]

Previous Estimates of the Extent of Testing Out

Two previous studies have provided estimates of the extent to which schools test out of program improvement. One study, which involved site visits at 27 school districts in 9 states, found that over half of the schools identified for program improvement tested out after 1 year, although "very few had initiated programmatic changes."[5] According to the other study, state Chapter 1 coordinators estimate that, on average, 50 percent of schools test out before implementing their improvement plans.[6] This study also reported that coordinators from states using a 0 NCE standard estimated, on average, a higher testing-out rate (60 percent) than those from states using higher NCE standards (40 percent).

PRINCIPAL FINDINGS

A High Percentage of Identified Schools Test Out

A substantial proportion of schools in our analysis that were initially below the NCE standard would have tested out of program improvement after 1 year. When we used a 0 NCE standard, three-fourths of the identified schools tested out; when we used a 4 NCE standard, about half of the identified schools tested out. (See fig. I.1.) The numbers in figure I.1 also confirm the conclusion of a previous study, cited above, that a smaller percentage of schools may test out when higher NCE standards are used.

---

[4]Mary Ann Millsap and others, The Chapter 1 Implementation Study: Interim Report, (Cambridge, Mass.: Abt Associates, Inc., under contract with the U.S. Department of Education, Office of Policy and Planning, 1992).

[5]Millsap and others, The Chapter 1 Implementation Study, p. 2-33.

[6]These officials' estimates varied considerably, however, ranging from 0 to 99 percent; 17 state coordinators said that less than one-third of their schools test out, and 15 said that more than two-thirds do so. Brenda J. Turnbull and others, Chapter 1 Under the 1988 Amendments: Implementation From the State Vantage Point, (Washington, D.C.: Policy Studies Associates, Inc., under contract with the U.S. Department of Education, Office of Policy and Planning, 1992).

Figure I.1: Percent of Identified Schools That Tested Out of Program Improvement After One Year

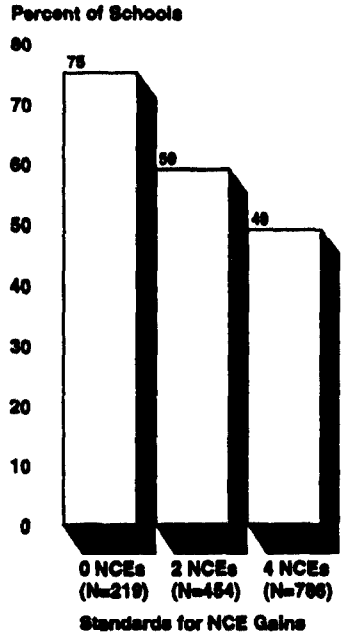


Figure reads: Of the 219 schools identified under the 0 NCE standard, 75 percent tested out the following year.

A variety of reasons may explain why schools that fall short of the NCE standard one year exceed it the next. In some cases, this may be attributed to partial implementation of an improvement plan. And even if improvement plans are not implemented, teachers may pay more attention to Chapter 1 instruction after a school is identified, which could lead to improved student performance on achievement tests. In other cases, however, a higher NCE change score might not represent real achievement gains. For example, schools could score below the NCE standard one year and above it the next simply because of the imprecision associated with NCE change scores. Although we do not know whether identified schools in our analysis implemented improvement plans, we do know that the imprecision of NCE change scores could have affected whether they were above or below the NCE standard in both years.

Most Schools Identified With Inconclusive Scores, Test Out With Conclusive Scores

Among the schools in our analysis that would have tested out of program improvement, a large majority would have been initially identified on the basis of inconclusive NCE change scores; that is,

their first NCE change score was not below the NCE standard by a statistically significant margin. In addition, a large majority of these schools would have tested out on the basis of conclusive NCE change scores; that is, their second NCE change score was above the standard by a statistically significant margin. (See fig. I.2.) Thus, for many of the schools that would have tested out, we cannot be highly confident that they should have been identified in the first place, but we can be highly confident that they exceeded the NCE standard after 1 year.

Figure I.2: Schools That Tested Out Usually Had Inconclusive NCE Change Scores in First Year, Conclusive Scores in Second Year

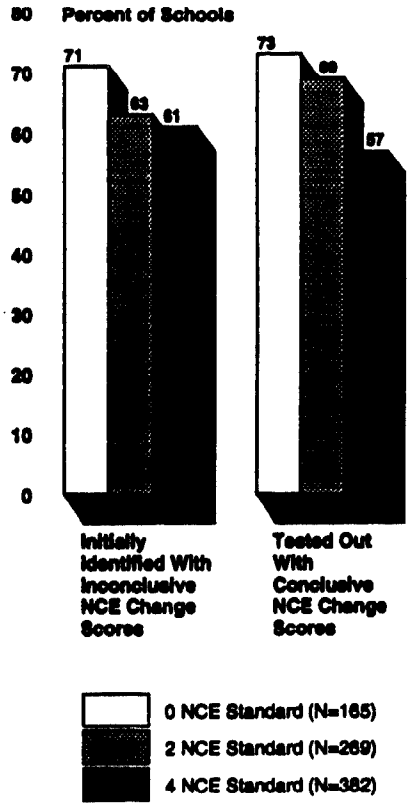


Figure reads: Of the 165 schools that tested out under the 0 NCE standard, 71 percent were initially identified with inconclusive NCE change scores.

ADDITIONAL TABLES SUPPORTING ANALYSIS OF TESTING OUT

The crosstabulations we computed as a basis for our analysis of testing out are presented in tables II.1, II.2, and II.3. These crosstabulations indicate whether schools' NCE change scores were above or below the NCE standard and whether these scores were conclusive (above or below the standard by a statistically significant margin) or inconclusive (above or below the standard by a statistically insignificant margin).

Table II.1: Relationship of Schools' NCE Change Scores to 0 NCE Standard Over 2 Successive Years

NCE change score in year 1	NCE change score in year 2				Total
	Above, conclusive	Above, inconclusive	Below, inconclusive	Below, conclusive	
Above, conclusive	875	178	59	42	1,154
Above, inconclusive	203	63	32	13	311
Below, inconclusive			19	11	147
Below, conclusive			7	17	72
Total	1,199	285	117	83	1,684

Note: Shaded area indicates schools that tested out.

Table reads: Of the 1,154 schools with a conclusive NCE change score above the 0 NCE standard in the first year, 875 were also conclusively above the standard in the second year.

Table II.2: Relationship of Schools' NCE Change Scores to 2 NCE Standard Over 2 Successive Years

NCE change score in year 1	NCE change score in year 2				
	Above, conclusive	Above, inconclusive	Below, inconclusive	Below, conclusive	Total
Above, conclusive	484	194	92	54	824
Above, inconclusive	221	96	51	38	406
Below, inconclusive			59	48	277
Below, conclusive			25	53	177
Total	890	374	227	193	1,684

Note: Shaded area indicates schools that tested out.

Table reads: Of the 824 schools with a conclusive NCE change score above the 2 NCE standard in the first year, 484 were also conclusively above the standard in the second year.

Table II.3: Relationship of Schools' NCE Change Scores to 4 NCE Standard Over 2 Successive Years

NCE change score in year 1	NCE change score in year 2				
	Above, conclusive	Above, inconclusive	Below, inconclusive	Below, conclusive	Total
Above, conclusive	220	126	85	61	492
Above, inconclusive	143	118	89	56	406
Below, inconclusive			83	77	393
Below, conclusive			60	184	393
Total	582	407	317	378	1,684

Note: Shaded area indicates schools that tested out.

Table reads: Of the 492 schools with a conclusive NCE change score above the 4 NCE standard in the first year, 220 were also conclusively above the standard in the second year.

All schools follow one of four patterns over a 2-year period, in terms of whether they are above or below the NCE standard. As shown in table II.4, below, schools that tested out represented a relatively small proportion of all Chapter 1 schools in our analysis. For example, with a 0 NCE standard, only 10 percent of schools were below the NCE standard one year and above it the next. An almost equal-sized group of schools (9 percent) followed the opposite pattern, moving from above the standard one year to below it the next. The most common pattern that schools followed over this 2-year period was to remain above the standard in both years. (See table II.4.)

Table II.4: Schools Follow One of Four Patterns Over a 2-Year Period

Patterns schools follow in relation to NCE standard over 2 successive years		Percentage of schools following each pattern, for various NCE standards (N=1,684)		
Year 1	Year 2	0 NCEs	2 NCEs	4 NCEs
Above standard	Above standard	78	59	36
Above standard	Below standard	9	14	17
Below standard	Above standard	10	16	23
Below standard	Below standard	3	11	24
	Total	100	100	100

SCOPE AND METHODOLOGY

DATA SOURCE AND SCOPE OF ANALYSIS

The state education agency in Pennsylvania provided us with two data sets, from all Chapter 1 schools in the state, containing Chapter 1 students' achievement-test scores in reading, math, or language arts. One data set had pretest scores from the 1988-89 school year and posttest scores from the 1989-90 school year; the other had pretests from the 1989-90 school year and posttests from the 1990-91 school year. These are the same data sets that Pennsylvania's state education agency uses to evaluate its Chapter 1 schools for program improvement.

We used several criteria to limit the scope of our analysis to include only certain types of test scores, students, and schools, as summarized below. Our analysis included only

-- achievement-test scores for reading, because far more schools had students with achievement-test scores for reading than for math or language arts;

-- test scores for advanced skills in reading, because we knew these were measured by students' scores on the reading comprehension portion of an achievement test, but we could not determine which subtest scores were used to measure basic reading skills;

-- students in grades 2 to 12, because schools are prohibited from using achievement-test scores for children below second grade for program improvement purposes;

-- students with pretest and posttest scores (hereafter "matched test scores") in a given school, because schools' NCE change scores are supposed to represent the average difference in the same students' achievement-test scores over a 1-year period (fall to fall or spring to spring); and

-- schools with matched test scores for more than 10 students, because schools that serve 10 or fewer students in Chapter 1 during an entire school year are exempted by law from consideration for program improvement.

Because of these limits on the scope of our analysis, our findings do not represent the actual number of Chapter 1 schools in Pennsylvania that were identified or tested out of program improvement during this time period.

METHODOLOGY

After applying the above criteria, and eliminating records with duplicate identification numbers and invalid test scores, we calculated an NCE change score for every school in each of the two initial data sets.

Next, we merged the two data sets, which produced a single data set containing 1,684 schools, each with two NCE change scores. Among these schools, the average NCE change score was 4.6 NCEs for the first year and 5.1 NCEs for the second year. The median number of Chapter 1 students per school in our analysis was 38 during the first year and 39 during the second. We used this merged data set to perform our statistical analysis.

To account for the imprecision of NCE change scores (described in enclosure I), we constructed confidence intervals at the 95 percent level around both NCE change scores for every school, based on a standard formula.[7] This means that we can be 95 percent confident that a school's "true" NCE change score falls between the upper and lower limits of the confidence interval. The fewer students with matched test scores a school has, the wider this range. For example, a school with matched scores for 15 students would have confidence intervals of + or - about 4 NCEs; a school with matched scores for 60 students would have confidence intervals of + or - about 2 NCEs.

Using these confidence intervals, we then classified both NCE change scores for each school into one of four categories: (1) above the standard by a statistically significant margin, (2) above the standard by a statistically insignificant margin, (3) below the standard by a statistically insignificant margin, and (4) below the standard by a statistically significant margin. We describe schools in the first and fourth categories as having "conclusive" NCE change scores and those in second and third categories as having "inconclusive" NCE change scores.

Finally, we cross-classified the data by year. This enabled us to determine not only the percentage of schools that exited program improvement 1 year after being identified, but also the percentage of schools that were identified the second year but not the first, the percentage not identified in either year, and the percentage

---

[7]To build confidence intervals, we first had to estimate the amount of overall variance in students' change scores due to measurement error. These calculations are described in greater detail in appendix III of our report Chapter 1 Accountability: Greater Focus on Program Goals Needed (GAO/HRD-93-69, Mar. 29, 1993).

identified in both years. This cross-classification also enabled us to determine whether the schools had inconclusive NCE change scores in one year, both years, or neither year. We did this analysis using three different standards for NCE gains (0 NCEs, 2 NCEs, and 4 NCEs) to reflect the range of NCE standards used in different states and districts around the country. (The three resulting crosstabulations are presented in enclosure II.)

(104746)