

Differences in test content and level of difficulty can radically affect the results shown by ostensibly similar tests and can even change the fundamental conclusions one reaches about the condition of educational achievement. For example, the apparent size of the achievement decline of the 1960s and 1970s--and even the presence or absence of a decline--varies with test content.

Even once the mix of skills and knowledge to be tested is determined, important decisions remain about the context in which the skills are to be assessed and the test's level of difficulty. For example, in the area of mathematics, the National Assessment of Educational Progress showed that the achievement decline of the 1970s was larger in the case of test items that embedded arithmetic skills in story problems than in the case of items that tested the same skills through simple computational exercises such as 23×45 . (Story problems are often seen as requiring higher-level skills--such as reasoning--in addition to rote computational skills.) The National Assessment also found no decline in the 1970s in lower-level reading skills (literal comprehension) but some decline in higher-level skills (inferential comprehension).

What Format Is Used?

Although the impact of test format--for example, multiple-choice, fill-in-the-blanks, open-ended short-answer, essay, and so on--is not completely understood, it is clear that format can affect the mix of skills actually tested and thus the results obtained.

In large-scale assessments, considerations of speed and cost create pressure to use a multiple-choice format. Multiple-choice tests can be graded quickly and unambiguously, often by machine. In contrast, scoring essay examinations can be time consuming, and guaranteeing even partial consistency among graders--or even among essays scored by a single grader--can be arduous.

Unfortunately, multiple-choice tests appear not to measure some higher-level skills well, though they can assess certain skills that are often referred to as higher level. For example, multiple-choice measures can test a student's ability to solve mathematical word problems, which require a higher level of skills than those required by simple computational exercises. Similarly, multiple-choice items can be designed to require sophisticated levels of reasoning, as a perusal of items from the SAT or ACT clearly

indicates. Nonetheless, research suggests that it is difficult--although not impossible--to write multiple-choice items that successfully measure certain aspects of reasoning, analytic thinking, and problem-solving abilities. As a result, performance on multiple-choice questions often depends more on factual knowledge and less on these higher-level skills than is intended. 9/

While this research indicates that multiple-choice tests have important limitations, it does not clarify the extent to which the use of such tests poses serious problems for the assessment of elementary and secondary school achievement. The degree to which the skills tapped by multiple-choice tests overlap with the set of skills that schools wish to foster remains a matter of debate but presumably varies considerably with subject matter and the age and ability level of students. Similarly, whether--or in what circumstances--the problems of alternative tests outweigh those of multiple-choice tests is a matter of argument.

How Well Does the Test Assess What It Is Intended to Test?

Whether achievement tests actually measure what they purport to is an underlying theme in the current debate about the proper role of testing.

Validity. The extent to which a test can be shown to test the skills that it is intended to test is called its *validity*. Simple subjective estimates of a test's validity are often misleading, and validity is therefore measured in a number of other ways.

In most cases, tests are validated by comparing performance on the test with some other criterion that can serve as a benchmark for the skills of interest. Unfortunately, straightforward criteria against which to validate achievement tests are rarely available. (If they were, the tests would often be superfluous.) For example, standardized tests originated in part as a substitute for teachers' judgments, which were deemed too subjective. Yet current standardized achievement tests are sometimes in part validated--for want of better criteria--by comparing scores on the tests with teachers' grades, as well as with scores on other similar tests. 10/

-
9. More discussion of this issue can be found in Norman Frederiksen, "The Real Test Bias: Influences of Testing on Teaching and Learning," *American Psychologist*, vol. 39 (March 1984), pp. 193-202.
 10. For example, see *SRA Achievement Series, Technical Report # 3* (Chicago: Science Research Associates, 1981).

One particularly important benchmark against which to validate tests is the closeness of the fit between the test and the curriculum to which students are exposed. This criterion--called *curricular validity*--has received increasing attention in recent years as a result of the spread of minimum-competency testing and the growth of litigation about test use.^{11/} If a test matches the curriculum poorly, it will provide misleading information about students' mastery of course material and about the effectiveness of teaching. It can also increase the influence that irrelevant factors--such as students' socio-economic background--have on scores and, in some cases, bias trends. ^{12/}

Reliability. Another characteristic of achievement tests that is closely tied to validity is the consistency of the scores they yield, which is referred to as *test reliability*. That is, if it were possible to administer equivalent tests several times, without the learning that would accompany repeated experience, how consistent would the results be from one administration of the test to the next? A reliable test is one that would show little variation; an unreliable test would show more. A test cannot be valid if it is highly unreliable, for the scores and rankings produced by an unreliable test largely reflect random error rather than the skills that the test purports to measure. It does not follow, however, that a test is valid merely because it is reliable; it can provide consistent estimates of the wrong thing. A highly consistent algebra test is not valid as a measure of knowledge of geometry.

-
11. For example, a central issue in *Debra P. vs. Turlington*--a suit concerning Florida's use of a minimum competency examination as a criterion for high-school graduation--was whether the skills and knowledge required by the MCT were actually taught in the Florida schools. *Debra P. et al., v. Turlington, et al.*, 474 F.Supp. 244 (U.S. Dist. Cr. Ct., Fla. 1979) Affirmed in part/Vacated in part/Remanded 644 F. 2d 397 (5th Cir. Ct. 1981).

Educators often draw a further distinction between curricular validity and instructional validity. The former refers to the correspondence between the test and the content of the curriculum materials, while the latter refers to correspondence with what is actually taught. (The courts have often spoken of curricular validity even when instructional validity was the principal issue.) While this distinction can be important in determining the validity of a test, it is not critical here, and both concepts are subsumed under the term "curricular validity" in this paper. See Peter W. Airasian and George F. Madaus, "Linking Testing and Instruction: Policy Issues," *Journal of Educational Measurement*, vol. 20 (Summer 1983), pp. 103-118.

12. For example, changes in curricular validity might underlie the fact that the ACT mathematics test results have not shown the sharp upturn that the SAT mathematics test results have shown in the past several years. Unlike the SAT, the ACT is intended to reflect the high school curriculum. One-fifth of the ACT mathematics test comprises geometry items, and a decline in the teaching of geometry as a distinct subject might be depressing scores, preventing an upturn like that of the SAT. (Personal communication, Mark Reckase, American College Testing Program, January 1985.)

Reliability is increased by repeated measurements. For example, a single measurement using an erratic thermometer would inspire little confidence, for a second reading might be very different. The average of many readings, however, would inspire more confidence, since the random errors would tend to be canceled out. Similarly, multiple measures of achievement are generally more reliable than a single measure. Indeed, adding additional information on a student's achievement will sometimes increase the reliability of the resulting conclusion even if the new information is itself less reliable than the old. For example, adding information about teachers' assessments of students to scores on a standardized test will sometimes increase the reliability of the conclusion even if the teachers' assessments are somewhat less reliable than the test. ^{13/}

All tests entail some unreliability, but that is generally not a problem when considering trends or comparison between groups, since the errors of measurement tend to cancel each other out when scores of many students are averaged. It can be a serious problem, however, when test scores are used to make decisions about individual students. Some of those decisions will invariably be incorrect if single tests are used as the basis for judgment. For example, consider a hypothetical requirement that students score above the average (475) on the SAT-mathematics to graduate from high school. About one-sixth of all students with "true" scores of 508 would obtain failing grades on any one administration of the test, as would about a third of students with true scores of 490. ^{14/} The SAT is widely considered to be a very well-constructed test, and the error rate using many other tests would likely be far higher.

How Are the Scores Scaled and Reported?

The scaling of test scores, and the form in which they are reported, can dramatically affect the results obtained, particularly when comparisons between groups or trends over time are of interest. Unfortunately, the ways of scaling and reporting scores that seem the most straightforward are often especially misleading.

-
13. Whether adding information from a less reliable measure increases or decreases reliability depends on the correlation between the various measures as well as the reliability of each. Adding information from a measure that is highly unreliable and largely uncorrelated with the original measure is more likely to reduce the reliability of the composite measure. Adding information from a measure that is nearly as reliable as the original and that is highly correlated with it is more likely to increase reliability.
 14. These calculations are based on a standard error of estimate of 34 points. Solomon Arbeiter, *Profiles: College-Bound Seniors, 1984* (New York: The College Board, 1984), p. iii.

One of the simplest methods of scoring tests is to express the scores as the percentage of items correctly answered, without regard for the relative difficulty of different items. This method is the standard in many classroom tests and was also the primary method of reporting results of the National Assessment of Educational Progress until recently.

Despite their outward simplicity, percentage-correct scores say relatively little about an individual's achievement and even less about the differences between individuals or groups. For example, what level of achievement would be indicated by a score of 50 percent correct on the National Assessment mathematics test? Is an improvement of 20 percentage points from that level comparable in significance to a decline of 20 percentage points? Lacking information about the level of difficulty of the items answered correctly or about the distribution of scores among students, these questions cannot be answered.

The most common solution to this problem is to translate scores into an alternative, comparative form that indicates where one student's score falls relative to all others. One common form is standard deviations, described earlier in this chapter. Another is percentiles. For example, the score of a student whose performance exceeded that of three-fourths of all others would be reported as being at the 75th percentile. Yet another, less commonly used now than in the past, is the "grade-equivalent score." In this scale, each student's score is expressed as the grade (often, year and month) of school in which the typical student attains a comparable score.

None of these scaling methods provides an unambiguous estimate of achievement differences between individual students or groups of students, but they can yield enough information to be useful. A comparative scale can indicate, for example, the percentile ranking that the average student in one ethnic group would attain if compared with students in another. It would not indicate, however, the relative amounts of skills and knowledge gained by typical students in both groups. A simple percent-correct measure provides less information. One can calculate, for example, the proportional difference between the average percent of correct answers in two ethnic groups (as has been done in Chapter 4 with the National Assessment data), but the meaning of those differences is unclear.

When comparing trends over time in different groups, the ambiguity of all of the scales becomes more serious. For example, consider a situation in which both low-achieving and high-achieving students appear to be gaining over time on a percentage-correct measure, but low achievers appear to be gaining faster. (A pattern of this sort appeared during part of the 1970s in some of the National Assessments.) For simplicity, say that the average

student in the low-scoring group went from having 20 percent to 40 percent correct answers, while the score of the average student in the high-achieving group increased from 80 percent to 90 percent. Without further information (such as the content and difficulty of the additional items each group answered correctly and the mix of items in the test), it is not obvious that the improvement in the lower group really reflects a greater achievement gain. For example, the improvement in the lower group might reflect a moderate increase in the proportion of many simple arithmetic items answered correctly, while the ostensibly smaller improvement in the higher group might reflect a sharp increase in the proportion of a few difficult algebra problems answered correctly. Information akin to this is rarely available from published sources, but even when it is, deciding which improvement is greater requires a subjective judgment. 15/

The use of comparative measures lessens these ambiguities, but it does not eliminate them. By using a comparative measure--such as standard deviations--one can ascertain which group changed more relative to the distribution of scores. Two ambiguities remain, however. First, the substantive meaning of a change from, say, 0 to 0.1 standard deviations (SDs) above the average might be quite different than that of an increase from 1.0 to 1.1 SDs above the average. On a mathematics test, for example, the first change might reflect improvements in computational abilities, while the second one reflected improvement in solving multi-step, multi-operation word problems. Second, different comparative measures can yield inconsistent answers. For example, relative trends expressed in SDs can be different from changes expressed in percentiles. In the previous example, an increase from 0 to 0.1 SDs above the average corresponds to an increase from the 50th to the 54th percentile, while the increase from 1.0 to 1.1 SDs above the mean--equivalent in terms of SDs--corresponds only to an increase from the 84th to the 86th percentile. Which of these measures is more meaningful is a matter of debate and depends in part on the question being addressed.

USING TESTS TO GAUGE TRENDS OR COMPARE JURISDICTIONS

The characteristics of the tests themselves are important in determining the results of achievement tests. But when tests are used to compare

-
15. The compression of high and low scores by percent-correct measures exacerbates this ambiguity. For example, in this instance, the high-achieving group could never show an improvement larger (in terms of simple differences) than that of the low-achieving group, for that would require scores above 100 percent correct.

jurisdictions (schools, districts, or states) or to gauge trends, several other considerations also become critical. These factors, while diverse, reflect a single underlying problem. In each case, the difficulty is that extraneous variation in test scores (for example, that reflecting disparities in students' backgrounds) is confounded with relevant variation (such as that attributable to differences in school effectiveness).

Differences in the Composition of the Tested Groups

Disparities in average test scores among jurisdictions need not indicate differences in the achievement of comparable students or, by implication, differences in the effectiveness of educational programs. Average test scores can differ, in some cases dramatically, because of disparities in the makeup of the groups of students tested. These compositional differences can have several sources.

One of the most important of these is differences in the ethnic composition of the student population. The gap in average scores between some ethnic groups tends to be very large, so even relatively small differences in ethnic composition can have a major impact on average scores. Moreover, differences in ethnic composition are often great. For example, the minority enrollments of the states varied in 1980 from 1 percent or less in Vermont and Maine to 57 percent in New Mexico, 75 percent in Hawaii, and 96 percent in the District of Columbia. Similarly, a 1982 survey of nearly 90 large school districts found minority enrollments ranging from over 90 percent in the District of Columbia, Atlanta, and Newark to 5 percent in Cobb County, Georgia, and Jordan County, Utah. 16/

Differences in dropout rates are another important source of compositional differences in the higher grades. Because dropouts tend to be low achievers, higher dropout rates will elevate a jurisdiction's average test scores.

Various educational policies also contribute to differences in the composition of tested groups. For example, rules governing the testing of handicapped students, the testing of students with limited proficiency in English, promotion from one grade to the next, and the testing of out-of-grade students can all have a substantial effect on average test scores.

16. CBO calculations based on data from the Office of Civil Rights, U.S. Department of Education.

All of these factors can bias trends as well as comparisons among jurisdictions in any one year. For example, districts experiencing atypically rapid growth in the share of their enrollments comprising certain minority groups would be likely to show less favorable trends than would others. Similarly, jurisdictions adopting particularly inclusive testing policies or finding successful methods to combat dropping out could make their achievement trends appear less favorable than they otherwise would. ^{17/}

How Are the Tests Made Comparable from Year to Year?

When trends in achievement are a concern, the methods used to make a test substantively comparable from year to year become critical in interpreting the results obtained. The simplest method of maintaining comparability over time is to keep the test the same. That is often unacceptable, however, for a number of reasons. Students and teachers might learn the content of a test, thereby artificially inflating scores--and lowering the test's validity--over time. Curricular changes might call for alteration of test content, and changes in student characteristics and performance might necessitate revision of test norms.

Faced with these problems, most test producers modify tests periodically and establish a new set of norms for the revised form. Scores on the revised test, however, need not be similar to those that the same students would receive if administered the old form.

In order to permit comparisons of the results of the old and revised forms, most test producers then estimate a mathematical relationship between the scores yielded by both versions. This process, called *equating*, can be done in several ways. The most straightforward is to administer both forms of the test to a single sample of students. In that case, differences in the scores yielded by the two versions must reflect changes in the test, and the scoring of the revised version can be adjusted to compensate, so that each student's score on the revised version is roughly that obtained on the old version. ^{18/} Another method requires including in the revised form a set of items from the old test. One can then administer

-
17. The impact of several compositional changes--such as changes in the self-selection of students to take college-admissions tests and trends in drop-out rates--on recent achievement trends is assessed in Congressional Budget Office, *Educational Achievement: Explanations and Implications of Recent Trends* (forthcoming).
 18. Because tests are not perfectly reliable, the scores obtained by an individual student on the two versions would not typically be identical even after this adjustment. Equating can remove much of the systematic change in scores attributable to revisions of the test, but other variation in students' scores remains.

the revised form to a sample of students and compare their scores on the new test as a whole with their scores on the shared items. If the relationship between performance on the shared items and scores on the old test in its entirety is understood, students' scores on the set of shared items can act as a proxy for the scores they would have received on the old test.

Annually Equated Tests. Annually equated tests are by far the most valuable in assessing achievement trends. When a test is equated every year, any given score reflects a comparable level of achievement in each year, and changes in scores can confidently be considered as differences in achievement. These differences, however, can reflect changes in the characteristics of the students tested as well as differences in the amount achieved by students of any given type.

Equating is a burdensome activity, and therefore very few tests are equated annually. In the absence of annual equating, interpretation of achievement trends is risky, although how risky depends on a variety of other aspects of the test. Accordingly, four tests that are annually equated --the SAT, the ACT, and the Iowa series of the Iowa Test of Basic Skills and the Iowa Test of Educational Development--are given particular attention in the analysis of trends in the following chapters.

Periodically Equated Tests. The periodic renorming of norm-referenced elementary and secondary achievement tests is the most common alternative to annual equating among tests that are formally equated at all. But it creates trend data that must be interpreted somewhat differently than are the data from annually equated tests.

Norm-referenced tests are typically renormed once every seven years or so, when new forms of the test are administered to national samples created by the tests' publishers. The resulting norms are used as a standard of comparison by schools that use the test for the following seven years or so. Publishers frequently equate the norming sample scores. This creates two types of information on trends: comparisons of norming-sample scores themselves, and annual comparisons of the scores obtained by districts and states using the test.

When test publishers equate the norming sample scores, comparisons of those scores can provide useful information on changes in achievement over the seven or so years between normings. Because each norming sample is intended to represent the national test-taking group at that time, the changes in the norms yielded by each sample in part reflect changes in the composition of the test-taking groups. The equating of norming sample scores, however, provides trend data that are in theory independent of changes in student characteristics.

These comparisons have two important limitations, however. First, because there are no comparable data from the years between normings, comparisons of norming sample scores can be misleading when achievement trends change over that interval. For example, if achievement was declining at the time of one norming but began increasing midway between then and the next norming, a comparison of the two norming samples might show no change at all--a pattern that would be entirely misleading unless annual data were available as a clue about trends in the intervening years. Second, in recent years questions have been raised about the adequacy of the publisher's national samples and changes in those samples over time stemming from changes in districts' willingness to participate in them.^{19/} Both nonrepresentativeness of norming samples and changes in their characteristics could substantially bias analysis of trends.

The annual, state- or district-wide data obtained from tests that are periodically renormed have a different set of advantages and disadvantages. During the period between normings--that is, while a single set of norms is used as the standard of comparison--these data provide a fairly good indicator of trends in the particular jurisdiction, except that growing familiarity with the test sometimes artificially increases scores or partially masks a decrease.^{20/} These trends, however, are confounded with changes in the composition of the test-taking group in the jurisdiction taking the test. On the other hand, during years of transition to a new set of norms, this system can produce serious distortions of achievement trends.^{21/} For

-
19. For example, Roger F. Baglin, "Does 'Nationally' Normed Really Mean Nationally?" *Journal of Educational Measurement*, vol. 18 (Summer 1981), pp. 97-108.
 20. Personal communication, Gene Guest, California Test Bureau of McGraw-Hill, December 1983.
 21. This distortion appears to have occurred, for example, in the Virginia statewide assessment, where adopting a new test form and set of norms produced sizable changes in scores in some subject areas that were not predicted on the basis of the national norming data. S. John Davis & R. L. Boyer, *Memorandum to Division Superintendents: Spring 1982 SRA Test Results* (Richmond: Virginia State Department of Education, July 19, 1982).

Periodically equated tests can also produce spurious changes when attempting to gauge a jurisdiction's level of achievement relative to the nation as a whole. For example, in a period when achievement is generally going up--as has been the case recently--most districts or states will see their scores rising relative to the old norms. This rise does not necessarily indicate that they are truly improving relative to the nation as a whole, but merely that the old norms are out of date. These jurisdictions are improving relative to what the national level of achievement used to be, but they could be improving either faster or slower than the nation as a whole.

this reason, the following chapters cite annual data from periodically normed tests only for the periods that a single set of norms was used.

Tests That Are Not Equated. Finally, some of the tests that have been used to illustrate recent achievement trends are not formally equated at all. The most important of these is the National Assessment of Educational Progress (NAEP), which was not equated until the most recent assessment of reading.^{22/} The absence of formal equating raises the level of uncertainty in any analysis of trends.

In the case of the NAEP, until recently the alternative to formal equating was to repeat a sizable proportion of the test items in subsequent assessments. Familiarity with test items is presumably not a problem in this case for a number of reasons: the test is administered only to a sample of children; it is administered only once every several years; and each student takes only a portion of the total test. Nonetheless, the procedure creates uncertainty. The method of assessing trends has most often been to compare adjacent assessments only in terms of the items shared by those assessments. The extent to which those items are representative, however, is open to question. Moreover, in at least one instance, the number of items shared over three assessments was so small that two different sets of items had to be used for the middle assessment--one for comparison to the earlier assessment (containing all items shared with that assessment), and another for comparison to the subsequent assessment.^{23/} This might have biased the assessment of trends.

Differences in Curricular Validity

Both analysis of trends and comparisons among jurisdictions can also be distorted by differences in curricular validity--that is, in the fit between a test and the curriculum. In both cases, the distortion is the same: groups for which curricular validity is lower will score comparatively lower than others, even if their actual level of achievement is similar. Typically, one might expect this problem to be less tractable when the domain of achievement being examined is complex than when it is narrow and simple. Devising a test of two-digit subtraction that has roughly comparable validity among districts, for example, might be much more feasible than designing

-
22. The most recent (1983) NAEP reading test was equated with all previous NAEP reading assessments (1970, 1974, and 1979).
 23. National Assessment of Educational Progress, *Three National Assessments of Science: Changes in Achievement, 1969-77* (Denver: NAEP/Education Commission of the States, 1978).

one in the area of intermediate algebra, which is broader and confronts designers of both curricula and tests with a wider array of choices.

The effects of curricular validity can be particularly vexing in assessing trends for another reason. When schools change the mix of skills they teach, there is no unambiguous way of equating tests over time unless some other criterion of achievement--independent of the schools' goals and curricula--is used as the basis for testing. For example, consider a situation in which an elementary school adds metric measurements to its mathematics curriculum, while eliminating the manual calculation of square roots. If a test that had high curricular validity before the change in curriculum is continued after the change, scores will decrease since students will more often fail to answer items about square roots, and there will be no items to compensate by testing their new knowledge of metric measures. 24/

One alternative is to change the tests to mirror changes in curriculum. If that is done, however, it is not obvious what levels of achievement are truly comparable among tests. Is proficiency in set terminology (a major addition to the mathematics curriculum during the years of the "new math") equivalent to facility in arithmetic computation (a mainstay of the "old" math)? While methods have been devised to estimate whether the items in the two domains are of comparable difficulty in a specific population, the question of whether these substantively different skills are "comparable" remains subjective. In addition, since changes in curriculum are generally only partly known, the question of whether the new and old tests have similar levels of curricular validity will remain in some doubt.

24. The effects of even relatively small changes in test content can be substantial, as is suggested by the recent experience of the statewide assessment program in Nevada, where changing to a revised form of the same norm-referenced test altered the ranking of districts in terms of average scores. This change in the districts' performance, however, might also reflect changes in test characteristics other than content--such as changes in format. (George Barnes, evaluation consultant, Nevada State Department of Education, personal communication, January 1985.)

CHAPTER III

AGGREGATE TRENDS IN EDUCATIONAL ACHIEVEMENT

Over the past several years, bad news has predominated in the public debate about educational achievement in the United States. Such developments as the decline in achievement that began in the 1960s, the unexceptional performance of American students relative to their counterparts in some other countries, and, most recently, the large gap in average achievement scores between black and white students have garnered widespread attention and have generated considerable concern. Less well known are some positive trends. For example, average achievement stopped declining some time ago and, by many measures, is rebounding sharply, and the gap between white and black students, while still large, has been shrinking.

THE DECLINE IN ACHIEVEMENT

Although not all indicators of educational achievement showed large declines over the past two decades, the great majority did, leaving no question that the decline was real and not an artifact of specific tests. The decline was widespread, appearing among many types of students, on many different types of tests, in many subject areas, and in all parts of the nation. Moreover, in many instances, the decline was large enough to be of serious educational concern.^{1/} Average scores declined markedly, for example, on the following achievement measures:^{2/}

-
1. The pervasiveness and magnitude of the decline were discussed in a number of earlier reviews. The breadth and size of the subsequent upturn in achievement, however, has not been previously assessed. Most of the early reviews were published before the characteristics of the upturn, or even its existence, were apparent. For earlier reviews of the decline, see especially Annegret Harnischfeger and David E. Wiley, *Achievement Test Score Decline: Do We Need to Worry?* (Chicago: ML-GROUP for Policy Studies in Education, 1975); also, Anne T. Cleary and Sam McCandless, *Summary of Score Changes (in Other Tests)* (New York: College Entrance Examination Board, 1976); and Brian K. Waters, *The Test Score Decline: A Review and Annotated Bibliography (Technical Memorandum 81-2)* (Washington, D.C.: Directorate for Accession Policy, Department of Defense, August 1981).
 2. See Appendix A for explanation of the principal data sources used in this paper.

- o College-admissions tests--the Scholastic Aptitude Test (SAT) and the American College Testing Program tests (ACT);
- o Most tests in the National Assessment of Educational Progress (NAEP);
- o Comparisons of periodic large representative samples of students --Project TALENT, the National Longitudinal Survey of the High School Class of 1972 (NLS), and the High School and Beyond (HSB) survey;
- o Periodic norming data from commercial standardized tests of elementary and secondary achievement;
- o The annual Iowa assessment of student achievement (which provides some of the most comprehensive and useful information on elementary and secondary achievement trends); 3/ and
- o A number of other state-level assessments of achievement.

On the other hand, a variety of achievement tests did not show large declines. In some cases, the exceptions were consistent over a number of tests, while in others, they appeared to be simply idiosyncratic. The most consistent exception was tests administered to children in the early elementary school grades. Among fourth-grade students, for example, declines appeared only inconsistently and were generally small. Moreover, there was apparently no substantial decline at all at even younger ages--by one measure, for example, third-grade scores showed a large, three-decade increase interrupted only by a brief pause and trivial decline in the 1960s and early 1970s. A variety of other tests--for example, the ACT natural science test--also showed only small declines or no decline at all. These exceptions, however, were so few that they do not call the overall decline into question.

When Did the Decline Begin and End?

The beginning of the achievement decline and its end showed markedly different patterns. To clarify the difference, it is helpful to distinguish between three patterns: "period effects," "cohort effects," and "age effects." In practice, a mixture of these three patterns is often found in achievement data.

-
3. The Iowa data are unique in providing annually equated data extending over many years, in many subject areas, and in all grades from 3 through 12 (see Appendix A).

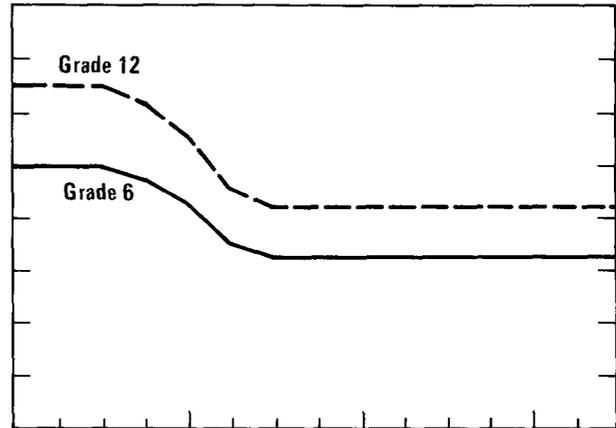
A period effect refers to a change that occurs in a specific time period, such as a decline in test scores that starts in roughly the same year among students of different ages or grade levels (see Figure III-1). In contrast, a cohort effect is a change that occurs with a specific birth cohort. An example would be a decline in scores that began with a particular birth cohort, appearing first in an early grade and then moving into the higher grades at a rate of roughly a grade per year as that birth cohort aged (see Figure III-1).

An age effect is a change that is linked to the age of those tested--perhaps occurring only in one age group, or varying in size from one age group to another. Age effects can occur with either cohort or period effects and, when data are incomplete, it can be impossible to disentangle them fully. For example, test scores have been rising in recent years. They started rising more recently in the higher grades, however, and to date have shown a smaller total increase in those grades than in the lower grades. This pattern could result entirely from the fact that scores in the higher grades have had fewer years to rise--that is, fewer of the cohorts contributing to the rise in scores have as yet reached the higher grades. In that case--a pure cohort effect--scores in the higher grades would be expected to continue rising in the near future as more of those cohorts pass through the higher grades (see Figure III-1). Alternatively, the pattern might reflect an age effect as well. Perhaps the lesser gains in the higher grades truly reflect less progress in those grades, as well as the later start of the upturn. This pattern might take the form of some cohorts not showing progress in the higher grades over the next few years comparable to that which they produced when in the lower grades (see Figure III-1).

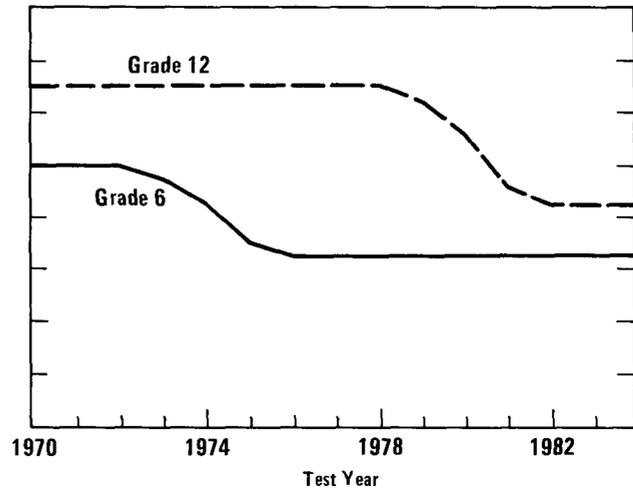
Very little information is available about the onset of the decline. Such information as there is suggests--albeit weakly--that the decline was a period effect, beginning relatively concurrently across a range of ages or grades. In contrast, the end of the decline--about which more data are available--shows a fairly clear cohort effect, occurring with a few specific cohorts of children and moving up through the grades as those cohorts passed through school.^{4/} On the other hand, given variation from test to

4. The period and cohort effects--if they are not an artifact of inadequate information--have substantial implications for the interpretation of the decline. Some observers have argued that period effects may be more consistent with the effects of changes in schooling, while cohort effects tend to suggest changes in student characteristics. See, for example, Christopher Jencks, "Declining Test Scores: An Assessment of Six Alternative Explanations," *Sociological Spectrum*, Premier Issue (December, 1980), pp. 1-15. This issue is discussed further in Congressional Budget Office, *Educational Achievement: Explanations and Implications of Recent Trends* (forthcoming).

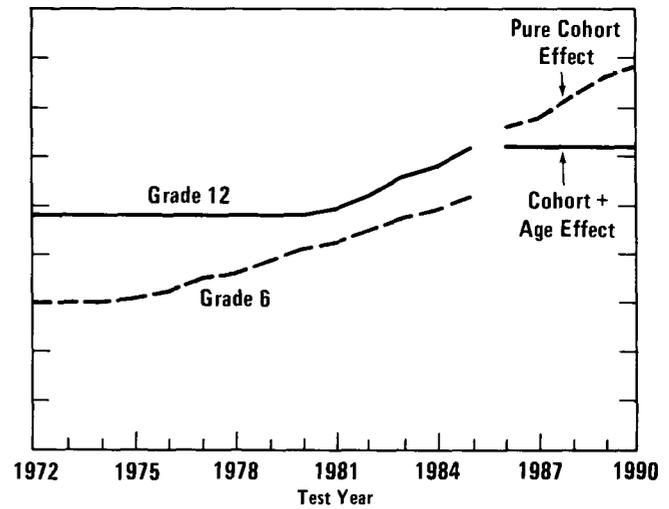
Figure III-1.
Hypothetical Period
Effect, Average Scores



Hypothetical Cohort
Effect, Average Scores



Hypothetical Age and
Cohort Effects,
Average Scores



SOURCE: Congressional Budget Office.

test and the paucity of data, the possibility remains that one or the other of these patterns--particularly, the period pattern shown by the onset of the decline--is merely a reflection of incomplete information. 5/

The few data sources that indicate the onset of the decline place it between the 1963 and 1968 school years (see Table III-1). The variation in the year of onset shows no obvious pattern from one test to another. The SAT began to decline in the 1963 school year. 6/ The decline in the ACT appears to have begun a few years later, in mid-decade. 7/ Scores in the Iowa statewide assessment--the Iowa Tests of Basic Skills (ITBS) through grade 8, and the Iowa Tests of Educational Development (ITED) in grades 9 and above--began dropping in every grade from 5 through 12 between 1966 and 1968. 8/ The Minnesota Scholastic Aptitude Test--a test independent of the College Board's SAT which was administered to high school juniors in Minnesota until the 1970s--began declining in 1967 after nearly a decade of uninterrupted increase. 9/

-
5. Only tests that provide annual or nearly annual data can be used to pinpoint the beginning and end of the decline. Many of the major data sources--such as the NAEP--have too great an interval between comparable tests to be useful in this regard.

Uncertainty about the timing of the decline's onset is heightened by the fact that the early decline on two of the four tests that can be used to pinpoint the onset--the SAT and ACT--was in substantial part a reflection of changes in the composition of the groups taking the tests. If there had been no such compositional changes, the timing of the decline on those tests might have been different.

6. Hunter M. Breland, *The SAT Score Decline: A Summary of Related Research* (New York: The College Board, 1976).
7. L. A. Munday, *Declining Admissions Test Scores* (Iowa City: The American College Testing Program, 1976). Scores on the ACT mathematics and social studies tests had already begun declining between 1964 and 1965--the first years of available data--but the decline was very small in the first year. The decline did not begin on the English test until 1966.
8. "Mean ITED Scores by Grade and Subtest for the State of Iowa: 1962-Present," and "Iowa Basic Skills Testing Program, Achievement Trends in Iowa:" 1955-1985 (Iowa Testing Programs: unpublished tabulations, 1984 and 1985).
9. Harnischfeger and Wiley, *Achievement Test Score Decline*.



TABLE III-1. ONSET AND END OF THE ACHIEVEMENT DECLINE, SELECTED TESTS

Test	Onset		End	
	Test Year	Birth Year	Test Year	Birth Year
SAT	1963	1946	1979	1962
ACT Composite	1966	1949	1975	1958
ITBS Grade 5	1966	1956	1974	1964
ITBS Grade 8	1966	1953	1976	1963
ITED Grade 12	1968	1951	1979	1962
Minnesota Scholastic Aptitude Test	1967	1951	N.A.	N.A.

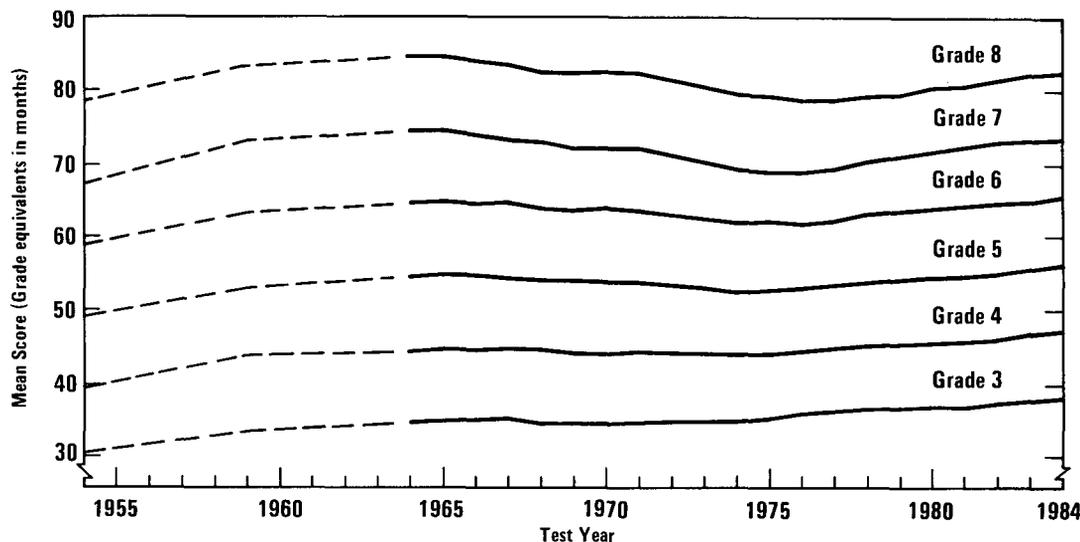
SOURCES: Hunter M. Breland, *The SAT Score Decline: A Summary of Related Research* (New York: The College Board, 1976), Table 1; *National College-Bound Seniors, 1985* (New York: The College Board, 1985); L. A. Munday, *Declining Admissions Test Scores* (Iowa City: American College Testing Program, 1976), Table 3; *National Trend Data for Students Who Take the ACT Assessment* (Iowa City: American College Testing Program, undated); Iowa Testing Programs, "Mean ITED Scores by Grade and Subtest for the State of Iowa: 1962-Present" and "Iowa Basic Skills Testing Program, Achievement Trends in Iowa: 1955-1985 (unpublished and undated); and Annegret Harnischfeger and David E. Wiley, *Achievement Test Score Decline: Do We Need to Worry?* (Chicago: ML-GROUP for Policy Studies in Education, 1975).

NOTE: N.A. designates not available.

The end of the decline (which can be ascertained with somewhat greater certainty because of more plentiful data) generally occurred within a few years of the birth cohorts of 1962 and 1963--that is, with the cohorts that entered school in 1968 and 1969. Thus, the low point in most achievement data occurred first in the lowest grades, moving into higher grades at a rate of roughly one grade per year as the cohorts of 1962 and 1963 passed through school.

This cohort pattern, which was first noted by those working with the Iowa tests (the ITBS and ITED), also occurs in a wide variety of other test

Figure III-2.
ITBS Composite Scores, Iowa Only, by Test Year
and Grade at Testing



SOURCE: "Iowa Basic Skills Testing Program, Achievement Trends in Iowa: 1955-1985" (Iowa Testing Programs, unpublished and undated material).

series.¹⁰ The progression through the grades is somewhat erratic--perhaps because of various unexplained year-to-year fluctuations in average scores--and is therefore not always apparent from a comparison of a few adjacent grades from a single test. The pattern becomes clearer, however, when a range of grades and tests are considered. Thus, the decline generally ended in the upper elementary grades in the mid-1970s, when the cohorts born within a few years of 1962 reached the ages of 10 and 11 (see Figure III-2). The decline in junior-high achievement ended a few years later. Tests given primarily to high school seniors (such as the SAT and the grade 12 ITED) stopped declining around the 1979 school year, when the birth cohort of 1962 was the appropriate age (see Figures III-3 and III-4).¹¹

10. Leonard Feldt, of the Iowa Testing Programs, the University of Iowa, pointed out the cohort pattern in the Iowa data (personal communication, December 1983).

This cohort pattern is particularly apparent in the Iowa data because they include annual information from all grade levels above grade three. In many other cases, the pattern becomes apparent only by comparing the timing of the decline's end among a variety of tests administered in different grade levels. See Appendix B.

11. One salient exception to this pattern is the ACT, which reached its low point a few years earlier.