assessments precludes drawing conclusions about the decline as a whole or placing the changes over the eight-year span into the context of longer-term trends in achievement. The time interval between the two studies includes, if other tests are an indication, a short period of rising scores as well as a longer period of declining scores. This mixture could distort assessments of the nature of the decline (particularly if the upturn does not parallel the decline in all respects) and could bias assessments of the impact of population changes on average scores.

IOWA TESTING PROGRAMS

Although many states have statewide testing programs, the data from the Iowa Testing Programs are uniquely valuable for the assessment of achievement trends. Unlike any other data source, it provides annually equated data extending over three decades for most grade levels in a variety of subject areas.

The Iowa data represent about 95 percent of public and private schools in the state. 28/ Unlike most statewide achievement data, the Iowa data do not reflect a mandatory, state-run program. Rather, they reflect voluntary participation by school districts in two testing programs administered by the University of Iowa. In grades 3 through 8, the test used is the Iowa Tests of Basic Skills (ITBS); in grades 9 through 12, it is the Iowa Tests of Educational Development (ITED). The ITBS is the same version as is administered in a large number of districts nationwide, while the ITED used in Iowa was a longer test than the version used elsewhere in the nation from the early 1970s until the most recent version. 29/ In both cases, the Iowa results are compared in this paper with statewide rather than national norms.

Both the ITBS and ITED tap a wide range of subject areas. The ITBS comprises 13 subtests in the areas of reading, vocabulary, language skills, mathematics, and work study skills. Trend data are available for all 13 subtests, but in most instances, only trends in a single composite score are reported in this paper. The ITED comprises seven tests: social studies, quantitative thinking, natural sciences, the interpretation of literary mater-

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 "Iowa Basic Skills Testing Program, Achievement Trends in Iowa: 1955-1985" (Iowa Testing Programs, unpublished and undated, 1985).

29. Robert Forsyth, Iowa Testing Programs, personal communication, March 1984.

ials, general vocabulary, correctness of expression (English usage), and sources of information (reference skills, knowledge of information sources, and so on).

The ITED is atypical of elementary and secondary standardized tests in that it includes no separate reading test. Instead, reading ability is assessed in the context of the substantive-area tests. Only in the last few years have the reading items from the various substantive-area tests been combined to provide a separate "reading total" score. Therefore, the "interpretation of literary materials" test, which taps many of the skills commonly included in reading tests, is used as a surrogate for a reading test in this paper, even though it is not a complete measure of the reading skills assessed by the ITED. 30/

The ITED is intentionally less closely tied to curricula than are some other standardized tests, although mastery of commonly taught materials is certainly necessary for success on it. The test aims to assess the intellectual skills that students will use in later life and those that represent the "long-run goals" of secondary schools. 31/ This intent is reflected, for example, in a very heavy emphasis on applications in the ITED quantitative thinking test. 32/

One major advantage of the Iowa data for assessing achievement trends is the length of the time span covered. Only the SAT provides data for a comparably long period. The Iowa data, however, have several additional advantages that the SAT does not share. The presence of data for 10 grade levels permits a clear assessment of the relationships between age and achievement trends and provides the single clearest test of the cohort pattern shown by the recent upturn in scores. The Iowa data also avoid two of the major problems of nonrepresentativeness inherent in college-admis-

30. For a summary of the content of the ITED tests, see *Iowa Tests of Educational Development, Forms X-7 and Y-7: Manual for Teacher, Counselors, and Examiners* (Iowa City: Iowa Testing Programs, 1979).

31. Iowa Testing Programs, *ITED Manual for Teachers, Counselors, and Examiners*.

32. Some of those working with the Iowa data believe that the much greater decline in mathematics scores shown by the grade-eight ITBS in comparison with the grade-nine ITED might reflect the fact that the ITED devotes more of its questions to applications and less to curriculum-based concept items than does the ITBS (Robert Forsyth, Iowa Testing Programs, personal communication, 1985).

sions test data: the Iowa data include students at all achievement levels and with all levels of educational aspirations. In addition, the Iowa tests, unlike college-admissions tests, are intended and designed to assess achievement rather than to predict subsequent college performance.

Nonetheless, the Iowa data have several important weaknesses for present purposes. Most important is the fact that Iowa is clearly not representative of the nation as a whole. For example, Iowa students on average score substantially above the national mean 33/. Moreover, minority students constitute a far smaller share of enrollments in Iowa than in the nation as a whole. 34/ Another limitation is that the available tabulations of the Iowa data include little information about the performance of important subgroups of students.

33. H. D. Hoover, Iowa Testing Programs, personal communication; Robert Forsyth, Iowa Testing Programs, note to school administrators (Iowa City: Iowa Testing Programs, unpublished, 1984).

34. As in the nation as a whole, however, minority enrollments have been increasing in Iowa. In 1972, minority students constituted 2.4 percent of enrollments in Iowa and 21.7 percent in the nation as a whole; in 1980, those proportions had grown to 4.1 percent and 26.7 percent, respectively (CBO tabulations of data from the Office of Civil Rights, U. S. Department of Education).

77-757 0 - 87 - 5 : QL 3

APPENDIX B

EVIDENCE OF A COHORT EFFECT IN THE RECENT UPTURN IN ACHIEVEMENT

Chapter III notes that the end of the achievement decline and the subsequent upturn conform more closely to a cohort pattern than to a period pattern. This Appendix provides more detailed data indicating the extent to which the trends conform to a cohort model. It has three sections:

- The first section explains the criteria that a data series must meet to provide a test of the models and identifies the best existing data for that purpose;
- The second section discusses the extent to which each of those data series is consistent with both models; and
- The final section pulls together data from a variety of series to provide a composite test of the models.

This Appendix is limited to the end of the decline and does not assess the extent to which the onset of the decline conforms to the period or cohort models. The data usable in assessing the characteristics of the onset of the decline are even more limited than those relevant to the decline's end. Thus, any characterization of the onset of the decline is largely speculative. 1/

TYPES OF DATA THAT CAN BE USED TO ASSESS COHORT AND PERIOD EFFECTS

Few of the existing data series on elementary and secondary achievement provide strong tests of the cohort and period models. To offer a strong test, a data series must:

1. Even some data series that extend back to the mid-1960s give no real indication of the timing of the decline's onset. Some of them (such as the social studies and mathematics tests in the ACT battery) were already declining at the time of the first available data. Moreover, two of the few test series with continuous data extending back into the 1960s --the SAT and the ACT--were seriously affected by major compositional changes in the test-taking population during the early years of the decline, leaving it unclear when they would have begun declining in the absence of compositional changes.

- Provide annual or nearly annual scores;
- Provide appropriate equating of scores, so that scores in one year can be considered comparable to those in other years;
- Extend over a period spanning at least one change in the direction of achievement trends (that is, one point at which average achievement stops increasing or stops decreasing); and
- Test reasonably comparable groups of students in different years. 2/

Further, the best test of the models is provided by data series that also provide similar measures of achievement at more than one grade level. Measures that are available only for a single age group--such as the SAT--provide a test of the cohort and period models only by comparing them with other tests that reflect different ages. Such a comparison can be biased by differences between the tests; the skills tapped by one test might show different trends than those tapped by another, and such a difference might be indistinguishable from a difference between cohorts or age groups. Few relevant data series, however, provide comparable measures in different age groups.

Within any single data series, the precise beginning of the decline or upturn is generally somewhat unclear, and therefore comparing several series is important. For example, the annual rate of change in test scores during the period around the end of the decline is typically very small, and average scores are therefore typically quite similar for a period of several years. This similarity introduces uncertainty into a choice of any year as the low point of the series and often makes it more meaningful to label a

2. The groups of students tested in each year need not be identical. Indeed, it is best if they are not identical in certain respects. But the confidence one can place in the data is lessened if the characteristics of those tested changed substantially more than the characteristics of the school-age population as a whole. For example, a sample that is entirely representative of the school-age population in each year would change over time (in terms of characteristics such as ethnicity, family structure, and poverty rates) as the school-age population changes. Such a sample would be optimal for testing the period and cohort models. On the other hand, compositional changes in the test-taking samples that are larger than those affecting the school-age population as a whole--such as those affecting the SAT candidate pool in the 1960s--can be sizable enough to mask period and cohort effects.

period of several years, rather than a single year, as the nadir. Comparison of a variety of series helps to lessen this uncertainty.

Given the criteria above, the following data series provide the strongest tests of the period and cohort models: 3/

- The Iowa Tests of Basic Skills, Iowa state series (ITBS-IA);
- The Iowa Tests of Educational Development, Iowa state series (ITED-IA);
- The American College Testing Program (ACT) college-admissions tests;
- The Scholastic Aptitude Test (SAT);
- The Virginia state assessment tests;
- The New York state assessment tests; and
- The California state assessment tests.

Two sources provide additional tests of the models, though they are weaker because they are not annual. One is the National Assessment of Educational Progress (NAEP). The second is the periodic renorming data from commercial standardized elementary and secondary tests. The latter are useful, however, only when publishers have retained data on equating studies contrasting the norms derived in each year.

THE FIT OF THE DATA WITH THE COHORT AND PERIOD MODELS

In this section, the fit of individual data series with the cohort and period models is examined. The patterns evident in the Iowa (ITBS and ITED) data are used as the point of comparison, since they provide the best single test of the models. The section first discusses data series that provide strong tests of the models, while those providing weaker tests (intermittent data, such as the NAEP) are left until the end of the section.

---

3. Additional detail on the characteristics of some of these data series can be found in Appendix A.

The Iowa Tests of Basic Skills, Iowa State Series (ITBS-IA)

The ITBS Iowa-state series reflects the scores of nearly all Iowa students through grade eight since the mid-1950s. In many respects, it is the best data on trends in elementary and junior-high achievement. Its advantages for the present purposes include:

- Equated data extending back to 1954, with annual data from 1964 to the present;
- Similar data on achievement in each grade through grade eight; and
- A general lack of problems with self-selection or other biasing selection changes in the student body taking the test.

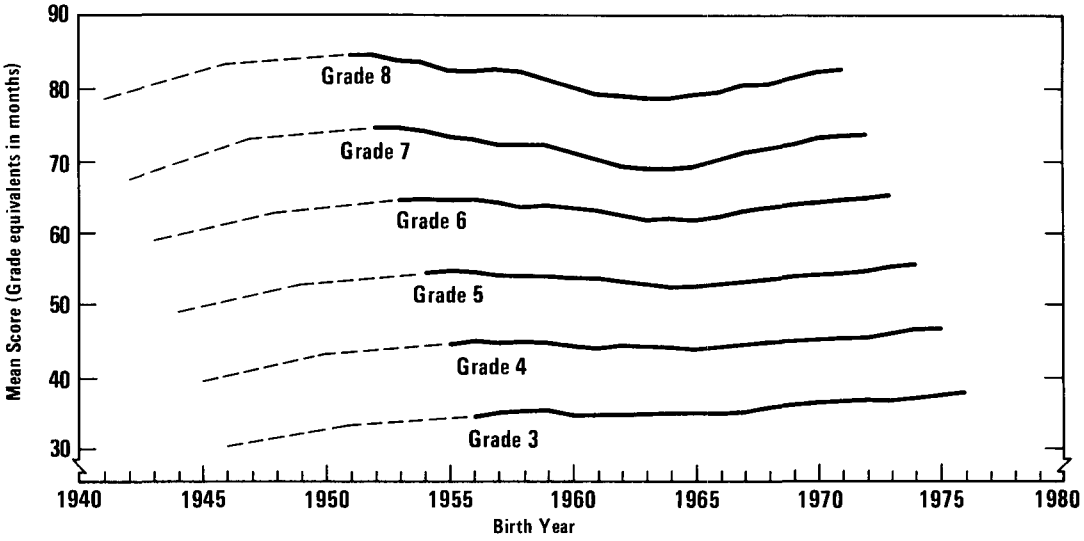
The greatest weakness of the ITBS-IA data is the fact that Iowa is in several important respects atypical of the United States. By some measures, average achievement in the elementary and secondary grades is nearly a grade higher in Iowa than in the nation as a whole. 4/ In addition, the student body in Iowa is demographically more homogeneous than the student body nationwide.

Average ITBS-IA scores reached their low points later in higher grades than in lower grades (see Figure III-2). Grade five scores bottomed out in 1974; grade six roughly in 1974; grade seven roughly in 1975; and grade eight in 1976. The changes in average scores in grades three and four are so small that it makes little sense to try to isolate a low point.

The later turnaround in higher grades suggests a cohort model, and the trends in grades five through eight indeed line up more closely when displayed in terms of birth years rather than year of testing (see Figure B-1). In grades seven and eight, the lowest scores reflect the birth cohorts of 1963 and 1964. The nadir occurred in grade six with the cohort of 1963, while in grade five it coincided roughly with the birth cohort of 1964.

4. H. D. Hoover, Iowa Testing Programs, personal communication, January 1984.

Figure B-1.
ITBS Composite Scores, Iowa Only (By birth year and grade at testing)



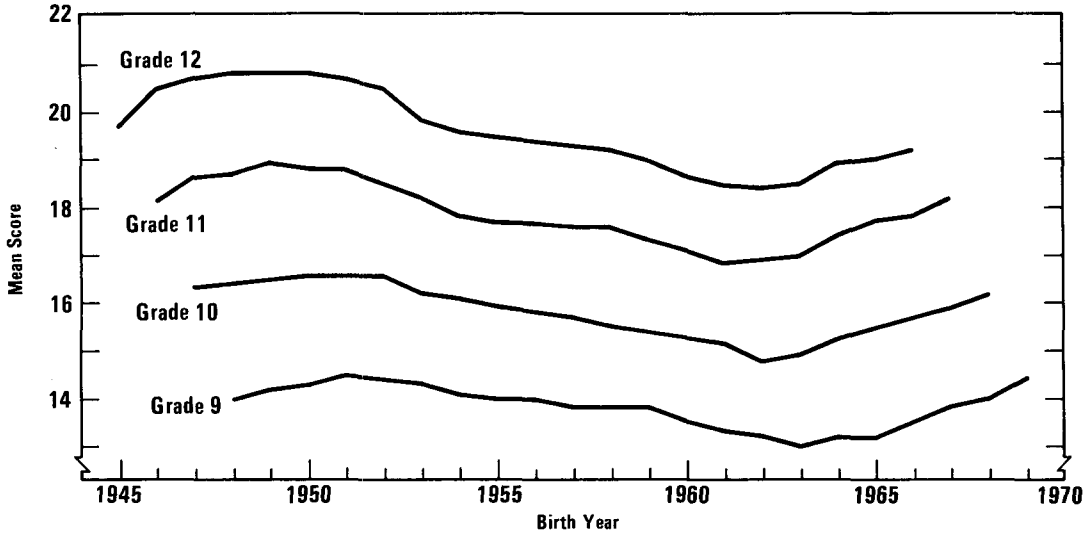
SOURCE: CBO calculations based on "Iowa Basic Skills Testing Program, Achievement Trends in Iowa: 1955-1985" (Iowa Testing Programs, unpublished and undated material).

The Iowa Tests of Educational Development, Iowa State Series (ITED-IA)

The ITED-IA, which includes grades nine through twelve, has the same strengths and weaknesses for the present purposes as does the ITBS-IA. Given the steeper achievement decline in the higher grades, the low point in the ITED is more clearly defined than that in the ITBS. The timing of the low points, however, provides less clear-cut evidence in favor of the cohort or period model.

When displayed in terms of test years, the ITED reached its low point in 1977 in grades 9 through 11, but not until 1979 in grade 12 (see Figure III-3). That is, grades 9 through 11 conform to a period model, while the entire span of grades 9 through 12 does not. Accordingly, when displayed in terms of birth years, the low points in the different grades do not fully line up (see Figure B-2). Grades 10 and 12 reached their low points with the 1962 birth cohort, while grade 9 was one cohort later and grade 11, one earlier.

Figure B-2.
ITED Composite Scores, Iowa Only (By birth year and grade at testing)



SOURCE: CBO calculations based on "Mean ITED Test Scores by Grade and Subtest for the State of Iowa: 1962 to Present" (Iowa Testing Programs, unpublished and undated tabulations).

If taken in the context of the ITBS results, however, the ITED trends can be seen as offering further support for the cohort model. Considering the two series together is logical, for while substantively the ITBS-IA and ITED-IA differ considerably, they largely reflect the same sample of students.

The earliest low point in the combined Iowa data occurred in 1974 in the grade five ITBS. The latest was in the ITED for grade 12, which reached its low point five school years later. The nadir in the junior-high scores occurred in between--roughly, in 1975 in grade seven, 1976 in grade eight, and 1977 in grade nine.

When tabulated in terms of birth cohorts, the low points in the combined Iowa data show less variation and less ordering from grade to grade. The earliest nadir was in the grade 11 ITED, which reached bottom

with the 1961 birth cohort. All of the remaining grades reached their low points with birth cohorts between 1962 and 1964. 5/

The Scholastic Aptitude Test (SAT)

The SAT data have the advantage of providing largely comparable scores from 1956 to the present. In addition, studies of the equating of SAT scores over time have been perhaps more extensive than those done with any other test. On the other hand, for present purposes, the SAT has several weaknesses:

- Serious problems with self-selection of students taking the test;
- Lack of comparable scores from a variety of grade levels; and
- Narrowness of the range of subjects covered (only two tests are administered--mathematics and verbal aptitude).

Enough is known about self-selection of students taking the SAT to know that those taking it are not representative of high school seniors in general. Not enough is known, however, to control fully for the non-representativeness of the SAT sample. On the other hand, while compositional changes--that is, changing self-selection--played a major role in the earlier (pre-1970) part of the decline in average SAT scores, they apparently have had only small effects in recent years. Moreover, they do not account for the turnaround in SAT scores, the timing of which is the most important aspect of the data for testing the cohort and period models. 6/

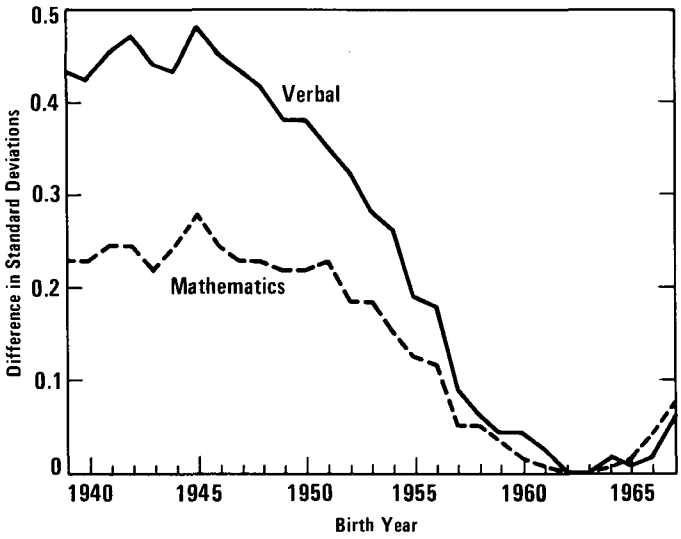
The end of the SAT decline fits the cohort pattern suggested by the Iowa data very closely. Both the mathematics and verbal scales of the SAT reached their minimums in the 1979-1980 school year, remained at that level for one more year, and then began their increases in the 1981 school year. Thus, the lowest scores reflect primarily the birth cohorts of 1962 and 1963, and the upturn began with the birth cohort of 1964 (see Figure B-3).

---

5. Grade six is ambiguous. It reached its low point somewhere between the birth cohorts of 1963 and 1965.

6. This point is discussed more fully in Congressional Budget Office, *Educational Achievement: Explanations and Implications of Recent Trends* (forthcoming).

Figure B-3.
Average SAT Scores (By birth year and subject; differences from lowest year)



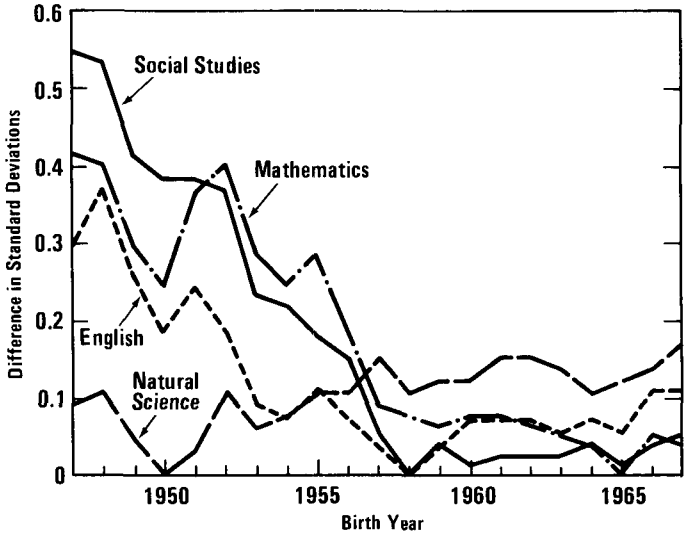
SOURCES: CBO calculations based on Hunter M. Breland, *The SAT Score Decline: A Summary of Related Research* (New York: The College Board, 1987); and The College Entrance Examination Board, *National College-Bound Seniors, 1985* (New York: The College Board, 1985).

The American College Testing Program (ACT) Tests

The ACT tests are also intended as college admissions tests, although they differ substantially from the SAT in format and content. The principal advantages and disadvantages of the ACT scores for present purposes are largely similar to those of the SAT. The ACT has the additional advantage, however, of covering a wider range of subjects: natural science and social studies, in addition to mathematics and English.

The end of the ACT decline is relatively clear-cut and is not consistent with the cohort pattern shown by the Iowa and SAT data. Average scores on the English and social studies tests bottomed out with the birth cohort of 1958, which was several cohorts earlier than those that

Figure B-4.
ACT Scores
(By birth year and subject; differences from lowest year)



SOURCE: CBO calculations based on L. A. Munday, *Declining Admissions Test Scores* (Iowa City: American College Testing Program, 1976), Table 3; and American College Testing Program, *National Trend Data for Students Who Take the ACT Assessment* (Iowa City: ACT, undated).

produced the lowest scores on the ITBS, ITED, or SAT (see Figure B-4). The mathematics trend is less clear. The major decline in scores ended with the birth cohort of 1959, but average scores moved down further, albeit slightly and erratically, until the 1965 birth cohort.

The ACT data also do not show the pronounced upturn in scores that characterizes the post-1963 birth cohorts in the SAT and Iowa data. Since the 1958 birth cohort, scores on the ACT test have fluctuated, showing only small and inconsistent increases (see Figure B-4). On the other hand, since the birth cohort of 1965--one to three years after the cohorts marking the bottom of the Iowa and SAT trends--the ACT tests have shown a fairly clear, but still very small, increase.

The New York State Assessment Data

New York State administers a wide range of tests to students of various ages, one of which provides a good test of the cohort and period models. In general, this one test conforms to the cohort model, showing timing that is largely consistent with that shown by the Iowa data and the SAT.

The Pupil Evaluation Program (PEP), begun in 1965, includes tests of reading and mathematics administered in grades three and six. Until recently, a norm-referenced test was used, and comparable annual data are available for spans of up to 16 years. Because the test is used to screen students requiring remedial services, the results are often tabulated in terms of the proportion of students falling below a threshold used for that purpose--the "state reference point." 7/

Three of the four tests--reading at both grade levels, and mathematics at grade six--conform to the cohort model suggested by the ITBS, the ITED, and the SAT. These three tests stopped declining with the birth cohort of 1962 and began improving markedly within a few years (see Figure B-5). Because the numbers are rounded and show no change for periods of two or three years before the upturn, the improvement might actually have begun with the cohorts a year or even two years earlier than 1963 or 1964, but that would still leave the timing consistent with the upturn suggested by the Iowa and SAT data. On the other hand, the proportion of students scoring above the reference point on the grade three mathematics test has been increasing almost without exception since the birth cohorts of the late 1950s. This exception is perhaps to be expected, however, given the general absence of sizable score declines in the earliest grades.

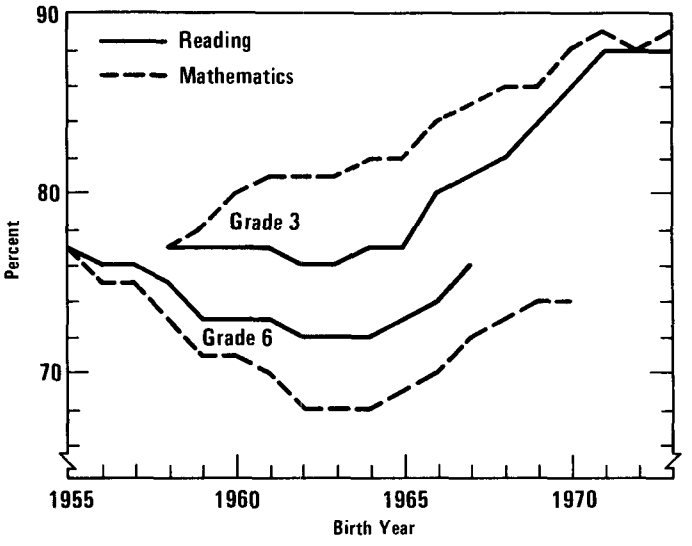
The California State Assessment Tests

Average scores of twelfth grade students in the California state assessment program fail to confirm either the cohort or period model, since they show very little change in any of the four subjects tested (see Figure B-6). The only appreciable year-to-year changes occurred between 1974 and 1975 (the birth cohorts of 1957 and 1958), and these changes were inconsistent in direction among subjects.

---

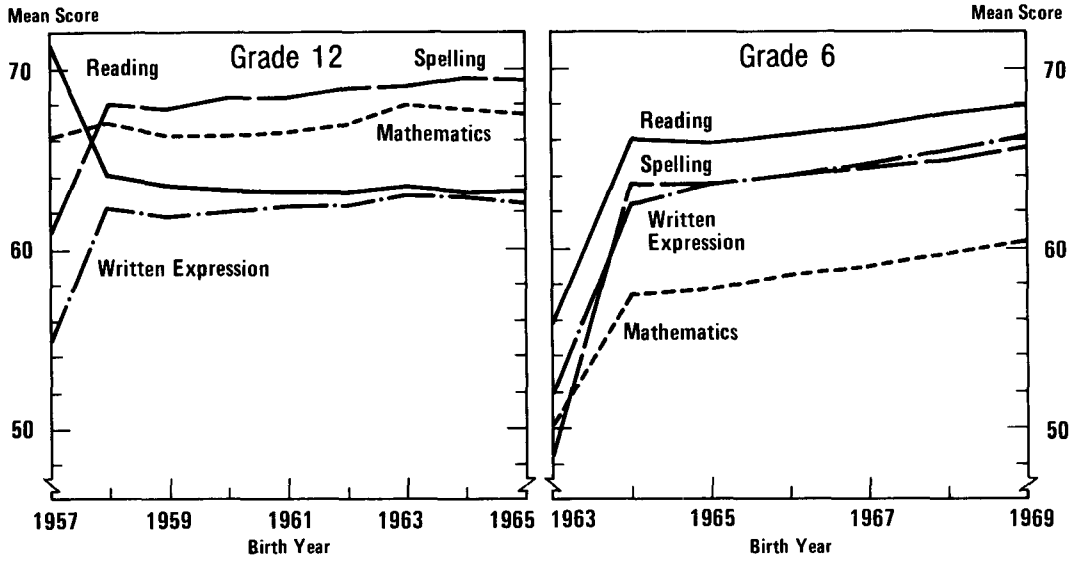
7. Division of Educational Testing, *Student Achievement in New York State 1982-83* (Albany: New York State Education Department, January 1984).

Figure B-5.
Percent of New York Students Scoring Above Reference Point (By birth year, grade, and subject)



SOURCE: CBO calculations based on Division of Educational Testing, *Percent of Pupils Scoring Below State Reference Point on Pupil Evaluation Program Tests* (Albany: New York State Education Department, undated).

Figure B-6.
California State Assessment Test Scores (By birth year, grade, and subject)



SOURCE: California Assessment Program, *California Assessment Program Summary Test Data* (Sacramento: California State Department of Education, undated).

Grade six scores from the California assessment also provide no support for either model (see Figure B-6). The birth cohort of 1964 scored substantially above the preceding cohort, but scores have risen only a small amount since then. Since the test was altered in the year that the 1964 cohort took the test (1975), this one-year increase in scores is likely to be a result of differences in tests rather than differences between cohorts.

The Virginia State Test Data

Data are available for the Virginia statewide assessment of fourth-, eighth-, and eleventh-grade students since 1972. During the seven-year period from 1974-1975 through 1980-1981, a single edition of one test (the 1971 edition of the SRA) was used. Because the same set of norms was used for scoring, the yearly averages from that time span can be compared with each other.

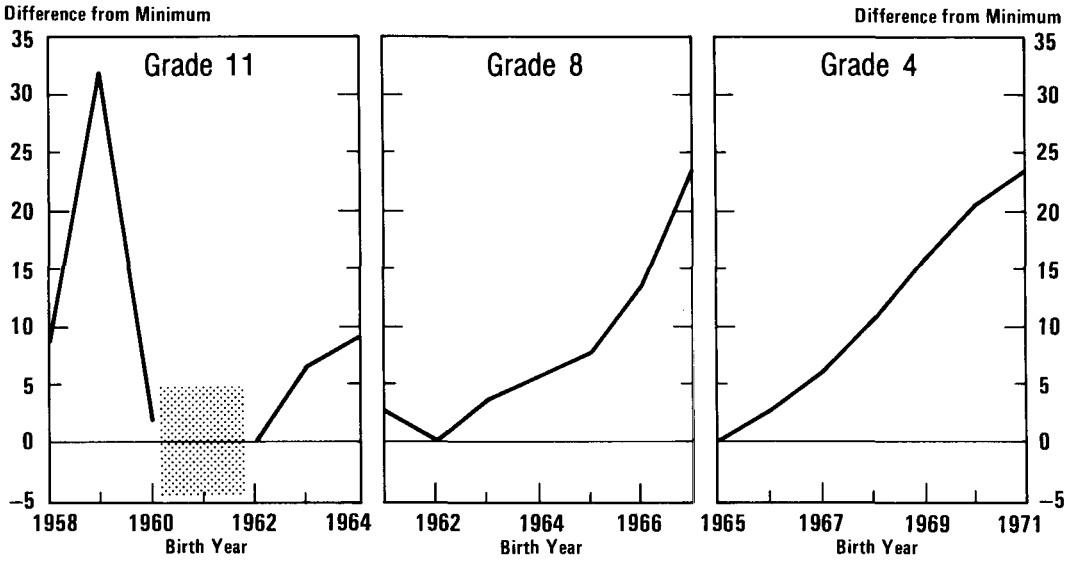
The Virginia assessment data provide a weaker test of the cohort and period models than do the data series above, but they provide a stronger test than do some of the intermittent data series discussed below. The relevant fourth-grade data begin only with the 1965 birth cohort, which is too recent to show the end of the decline if the cohort model is correct. The eighth grade data do span the end of the decline, but only barely; the first data point is the 1961 birth cohort. The eleventh grade data span the end of the decline nicely but lack information for the birth cohort of 1961.

Given these limitations, the composite scores from the Virginia data appear to conform closely to the cohort model (see Figure B-7). Among eleventh graders, the low point appears to have occurred with the birth cohorts of 1961 or 1962, although the large increase between the 1958 and 1959 birth cohorts calls the stability of the scores into question. The average scores of eighth graders appears to have reached its low point with the 1962 birth cohort, though the absence of data before the 1961 cohort leaves some doubt about that. Finally, fourth-grade scores have been increasing from the first year of data, which is consistent with the cohort model. Since the earliest data are for the 1965 cohort, however, this fact offers the model only weak support. Scores on the specific subject-area tests that enter into the composite scores (reading, mathematics, and science) show largely similar trends, except that the upturn among eighth graders is less clear-cut in reading.

The National Assessment of Educational Progress (NAEP)

The NAEP data reflect assessments at intervals of up to five years. As a result, they provide only a weak test of the cohort and period models. They cannot pinpoint the year in which the decline ended or even confirm that

Figure B-7.
Virginia Composite Achievement (By birth year and grade)

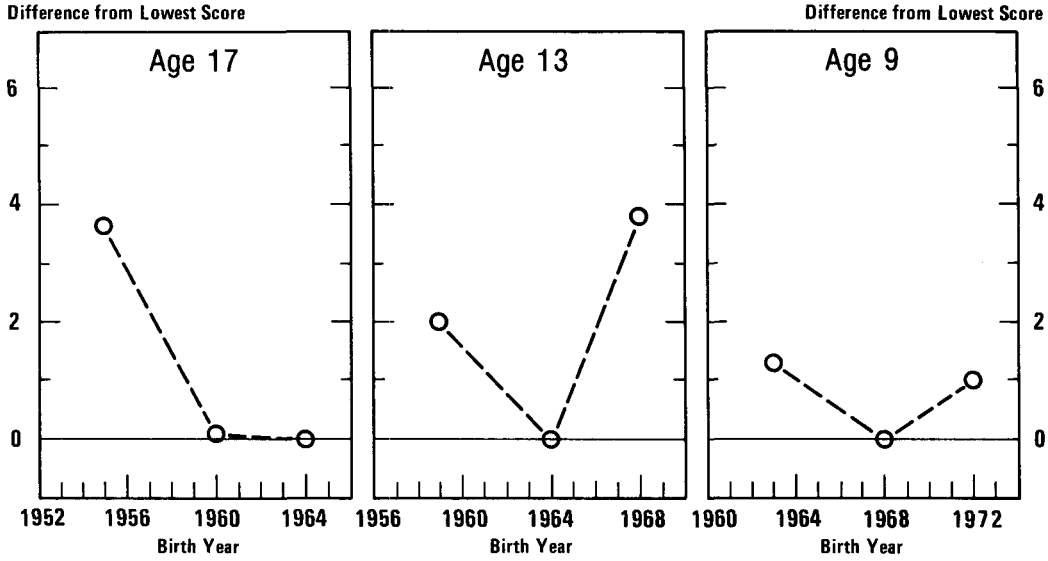


SOURCE: CBO calculations based on S. John Davis and R. L. Boyer, *Memorandum to Division Superintendents: State Testing Program Results, 1980-1981* (Richmond: Commonwealth of Virginia Department of Education, 1981).

there was only one recent change in the direction of the trend--that is, only one recent period each of decline and upturn.

For example, the NAEP mathematics scores of 13-year-olds reached their lowest recorded average with the assessment of 1977--that is, with the birth cohort of 1964 (see Figure B-8). The true low point, however--assuming that there was only one--might have occurred with any of the birth cohorts from 1960 through 1967. For the low point to have occurred within a few years of the tested cohorts of 1959 or 1968 is unlikely, for that would have required very abrupt changes in average scores, but a considerable range of alternatives to the apparent low of 1964 remain plausible.

Figure B-8.
NAEP Mathematics Scores (By birth year and age)



SOURCE: CBO calculations based on National Assessment of Educational Progress, *The Third National Mathematics Assessment: Results, Trends, and Issues* (Denver: NAEP/Education Commission of the States, 1983).

Moreover, the NAEP data are not entirely consistent--even within these limits--with either the cohort or the period model. On balance, the data seem more consistent with a cohort model and suggest an upturn that began, as in the SAT, Iowa, Virginia, and New York data, with the birth cohorts of the first half of the 1960s. There are enough exceptions, however, that some observers might disagree with this generalization.

Of the NAEP data, the mathematics results are least consistent with a cohort model and, conversely, most supportive of a period interpretation (see Figure B-8). In the case of both 9- and 13-year olds, the lowest average score occurred in the 1977 assessment--that is, with the birth cohorts of 1968 and 1964, respectively. This pattern is entirely consistent with a period model. The actual lowest points, however, might have occurred in years when there was no assessment and thus might differ

between the two age groups. In the case of 17-year-olds, the low point was marked by both the 1977 and 1981 assessments, since the average scores in those two years were effectively equal. On the other hand, the data from the 13- and 17-year-old groups--but not that from the 9-year-olds--is also consistent with a cohort model. If the cohort model pertains, these data suggest that the minimum occurred with the birth cohorts of the first half of the 1960s--perhaps, in the range of 1961 through 1965.

The NAEP science and reading assessments are somewhat more supportive of the cohort model, although in these subjects also the patterns are not clear-cut. The science data, regardless of age, provide no indication of further sizable drops after the birth cohort of 1963, although the absence of comparable tabulations from the most recent assessment calls this into doubt and leaves open the possibility of a period effect (see Figure B-9). The NAEP assessments never showed a sizable decline for reading as a whole, but the reading data do suggest that average achievement began rising with the birth cohorts of the early 1960s or late 1950s (see Figure B-10). (The scores of 13-year-olds are in this case a rare exception in suggesting the possibility of an upturn that began before the cohorts of the 1960s.) The NAEP assessment of inferential comprehension in reading--which, unlike the data for reading as a whole, did show a decline--also is consistent with the view that the decline ended and the upturn began with the cohorts of the early 1960s (see Figure B-11).

The ITBS National Norming Data

The ITBS, like most commercial standardized elementary and secondary tests, is renormed approximately once every seven years. The ITBS norming data reported here, unlike the ITBS-IA data described above, is based on national samples of students. 8/

8. Although norming data need not be useful in assessing national trends in test scores, the norming of the ITBS and certain other tests does yield valuable information on trends. The principal purpose of renorming is to estimate the national distribution of scores on a new version of the test, so that districts using the test have an updated national standard against which to judge their own scores. This objective does not necessitate equating the old and new versions of the test. The two versions often are equated, however, and the results of the equating provide an estimate of the change in the national distribution of scores. All ITBS norming results have been equated to previous norming-sample results.

   Equated national norming data are available for the ITED as well but are not discussed here. The ITED averages declined between the two most recent normings (1971 and 1978), but there has been no renorming since then. As a result, there is as yet no evidence of the overall upturn in scores. Lacking that, the ITED norming data provide no information on the timing of the decline's end.