DOCUMENT RESUME

ED 284 941 UD 025 688

TITLE Trends in Educational Achievement. CBO Study.
INSTITUTION Congress of the U.S., Washington, D.C. Congressional Budget Office.
REPORT NO 59-115-0-86-1
PUB DATE Apr 86
NOTE 178p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC08 Plus Postage.
DESCRIPTORS *Academic Achievement; *Achievement Rating; *Achievement Tests; Blacks; *Educational Testing; Elementary Secondary Education; Hispanic Americans; *Minority Groups; Testing; Testing Problems; *Test Interpretation; Test Use; Urban Schools

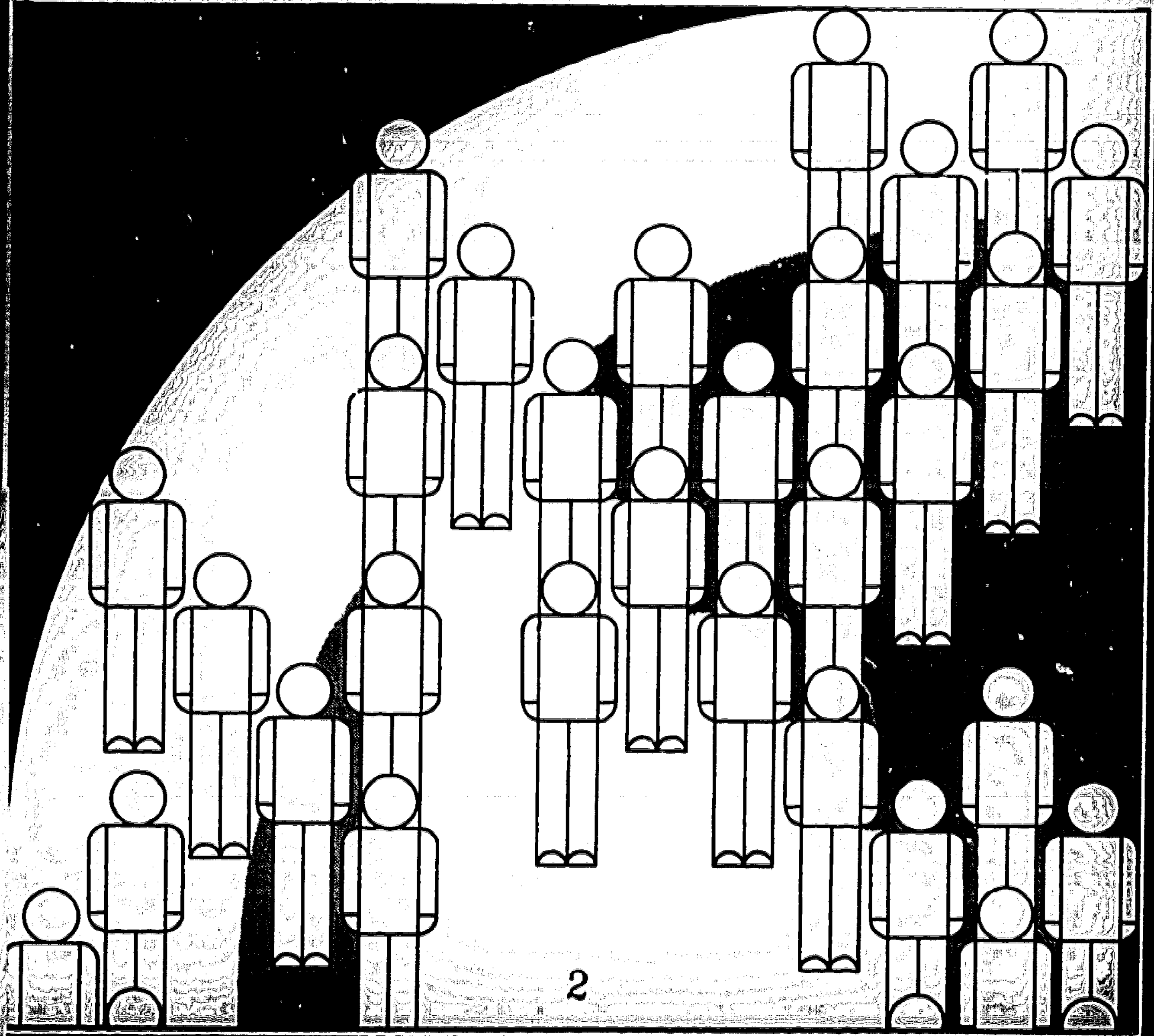
ABSTRACT

This publication presents assessments of trends in the educational achievement of elementary and secondary school students. In light of the heightened reliance on achievement tests a careful appraisal of recent trends in scores has important ramifications for educational policy. This study assesses test score trends and offers some insights on the strengths and weaknesses of the information they provide. While the 1960s saw a decline in achievement scores in grades five and above the decline was primarily in areas involving higher order skills rather than basic skills. The test score decline ended in the upper elementary grades beginning in the mid-1970s. Achievement has been steadily rising; however, examination of the data raises qustions as to whether these score improvements on some tests have been larger in the more basic skills areas than in areas requiring deeper understanding. Minority students' performance on tests has improved over the past 10 to 15 years and the gap between black and white students' scores has narrowed. Further, Hispanic students have also made gains over the past decade, with the greatest improvement being among Mexican Americans. Finally, scores have improved in characteristically low-achieving urban schools and communities. These and other findings are discussed in detail throughout the body of the report and are supported by extensive statistical tables and charts. (CG)

CONGRESS OF THE UNITED STATES
CONGRESSIONAL BUDGET OFFICE

APRIL 1986

Trends in Educational Achievement



CBO STUDY

TRENDS IN EDUCATIONAL ACHIEVEMENT

The Congress of the United States
Congressional Budget Office

NOTES

Except where otherwise noted, dates used in this paper are school years rather than calendar years. For example, the results of a test administered in the fall of 1979 and the spring of 1980 are both labeled 1979. As a result, the dates used here are in some instances a year earlier than those in other published sources. This discrepancy is particularly common in the case of college admissions tests and other tests administered to high school seniors, which are often labeled in other sources in terms of the calendar year in which students would graduate.

Details in the text and tables of this report may not add to totals because of rounding.

PREFACE

At the request of the Subcommittee on Education, Arts, and Humanities of the Senate Committee on Labor and Human Resources, the Congressional Budget Office (CBO) prepared this assessment of trends in the educational achievement of elementary- and secondary-school students. This volume presents the analysis of the trends themselves; a forthcoming companion volume, *Educational Achievement: Explanations and Implications of Recent Trends,* evaluates many common explanations of the trends and discusses their implications for education policy. In accordance with CBO's mandate to provide objective and impartial analysis, neither volume contains recommendations.

Daniel Koretz of CBO's Human Resources and Community Development Division prepared the analysis under the direction of Nancy M. Gordon and Martin D. Levine. Paul L. Houts edited the report. Ronald Moore typed the many drafts and prepared the manuscript for publication.

The author thanks the following organizations for providing essential data, much of which is unpublished: the National Assessment of Educational Progress, the American College Testing Program, the Iowa Testing Programs, the College Board, CTB-McGraw Hill, Science Research Associates, the state departments of education in New York, Virginia, Texas, North Carolina, Nevada, California, and Illinois, and the school districts of Cleveland, Montgomery County (Maryland), and Houston.

Many individuals contributed in various ways to this work. Particular thanks are due H. D. Hoover of the Iowa Testing Program and Lawrence Rudner of the Office of Educational Research and Improvement, U.S. Department of Education, who provided insightful contributions at many stages of the project. Other individuals whose assistance is gratefully acknowledged include Robert Cameron and Harlan Hanson of the College Board; Douglas Coulson of the Office of Technology Assessment, U.S. Congress; Robert Forsyth of the Iowa Testing Programs; Gene Guest of CTB/McGraw-Hill; Eric Hanushek of the University of Rochester; Lyle Jones of the University of North Carolina; Jackie Woods of the American College Testing Program; and Edward M. Gramlich, Jack Rodgers, and Roberton Williams of CBO. Kenneth Rubin of CBO also provided particularly valuable comments.

CONTENTS

SUMMARY

Over the past several years, the educational achievement of American students has become a focus of intense public discussion and has led to a serious reexamination of schooling in America. A number of developments have contributed to this concern, including a substantial decline in test scores in the 1960s and 1970s, the weak performance of American students relative to their peers in some other countries, and the large gap in average test scores between some minority groups and nonminority students. More positive trends, though significant, have gained less notice--for example, the end of the overall achievement decline in the 1970s, a subsequent upturn in average scores, and recent gains of black and Hispanic students relative to nonminority students.

With the growing concern about public education has come an increasing reliance on achievement tests as indicators of the performance of students and schools. This trend has taken many forms and is apparent from the local to the national level. Many states and localities have expanded their programs of routine testing, sometimes as a result of legislation; the additional tests are often used as minimum criteria for promotion into higher grades or for graduation. Furthermore, average test scores have become a common basis of comparisons among schools and districts, and in some communities, newspapers routinely publish test results to facilitate such comparisons. The U.S. Department of Education has begun annual publication of average college admissions test scores on a state-by-state basis, and some states have taken steps to alter their own achievement tests to make their results comparable. Test scores have in fact come to be used as a national report card, influencing decisions from the level of individual students to that of national educational policy.

In the light of this heightened reliance on achievement tests, a careful appraisal of recent trends in test scores has important ramifications for educational policy. This paper assesses test score trends among elementary and secondary school students; it also discusses the strengths and weaknesses of the information they provide. A forthcoming companion study, *Educational Achievement: Explanations and Implications of Recent Trends*, evaluates common explanations of the trends and explores implications for educational policy.

THE POLICY CONTEXT OF CURRENT CONCERNS

Although states and localities bear primary responsibility for elementary and secondary education, educational achievement is clearly a national concern. Indeed, the current debate has been national in both scope and content. It has focused in part on such national issues as the competitiveness of the American economy and national security--questions that have been recurrent themes in debate about educational policy at least since the turn of the century. Moreover, the debate has taken hold in all regions of the country, and many of the initiatives undertaken by states and localities reflect common themes and share common elements, such as increased reliance on achievement testing. As in the past, both the Congress and the Administration have been important participants in the debate through legislative proposals and the dissemination of information.

UNDERSTANDING MEASURES OF EDUCATIONAL ACHIEVEMENT

Although the use of standardized tests as indicators of educational achievement has grown sharply in recent years, scores on standardized tests are not as straightforward an indicator of achievement as they might first appear. For that reason, the strengths and weaknesses of existing tests should be kept in mind when interpreting recent trends.

The advantages of standardized tests--or, rather, the advantages that they can have if carefully constructed--are obvious and important. By imposing a uniform measure, they can avoid much of the subjectivity and extraneous variation that plagues some alternative forms of evaluation, such as grade-point averages. Standardized tests can be designed to provide valuable comparisons over time and among grade levels, tap specific types of skills, and differentiate among students at various achievement levels.

The weaknesses of standardized tests are less apparent but equally significant. In most cases, the tests are not direct and complete measures of the skills that are of concern. Rather, they are proxies for this often unobtainable ideal. Designing the proxy entails many decisions about the test's purpose, content, level of difficulty, format, the severity of time pressure, and other factors. As a result, tests vary markedly in what they measure and how well they measure it. Indeed, even apparently similar tests often produce divergent results.

Tests designed to assist in selecting students for admission to college--such as the Scholastic Aptitude Test (SAT)--provide a particularly striking example of tests as proxies for other, unobtainable measures. These tests are intended to predict students' performance in college, which can be measured directly only long after the admissions decision must be made. Although these tests comprise multiple-choice questions, their purpose is to predict future success on some very different tasks--such as comprehending long lectures and writing fluent term papers--that help determine whether students succeed or fail in college. In the case of tests designed to measure students' current level of achievement, the contrast between the skills embodied in a test and the corresponding skills with which schools are concerned is often less striking, but it can nonetheless be substantial.

Because of these limitations, the results of standardized tests must be interpreted cautiously. Trends should be given credence if they appear with considerable consistency in numerous tests, particularly if the tests are varied. On the other hand, trends that appear only on one test, or only among a set of very similar tests, should be considered questionable. Moreover, whether trends shown by a test are meaningful hinges on whether the characteristics of that test are appropriate for the particular issue in question. For example, if trends among students in general are at issue, college admissions tests can provide dubious information. A large number of students never take such tests, which makes the results unrepresentative of the student population as a whole. Furthermore, biases are introduced by changes in the composition of the group that does take the tests. Similarly, some minimum-competency tests provide little information about trends among high-achieving students for want of a sufficient number of difficult test items.

THE DECLINE AND SUBSEQUENT UPTURN IN ACHIEVEMENT TEST SCORES

After years of improvement, scores on achievement test scores began a sizable drop in the mid-1960s. The decline was widespread, occurring among many different types of students, on many different tests, in all subject areas, in private as well as public schools, and in all parts of the nation. [1]

Although the size of the decline varied greatly from one test to another, it was in many instances large enough to be of substantial educational concern. In general, the decline in test scores was larger in the

[1] A few tests did not conform to this pattern. The National Assessment of Educational Progress (NAEP), for example, showed no overall drop in reading since 1970, and the American College Testing program (ACT) tests showed no decline in natural science. But these exceptions were few enough, and the conforming tests sufficiently numerous, that the generality of the decline is clear.

higher grades. Scores on tests administered in grades three and below dropped little, if at all, and tests administered in grade four showed only inconsistent and small declines. On the other hand, most tests administered in grades five and above showed declines in average scores, with the largest drops tending to occur at the high school level. Among the achievement tests assessed in this study, the average decline in grades six and above was large enough that the typical (median) student at the end of the decline exhibited the same level of achievement as was shown before the decline by students at the 38th percentile.[2] A different assortment of tests, however, would yield a different estimate of the decline's average magnitude.

Although not all skills commonly considered "basic" escaped serious deterioration, the score decline appears to have been greater in areas involving higher-order skills. For example, between 1972 and 1977, the National Assessment of Educational Progress in mathematics showed no change in the performance of 17-year-olds in the simple recall of facts and definitions, but substantial declines took place on test items tapping deeper understanding and problem-solving skills. Items testing arithmetic computation showed a mixed pattern; in general, the more complex items evidenced the sharpest drops in success rates. This larger drop in higher-level skills might be one cause of the greater test score decline in the higher test grades.

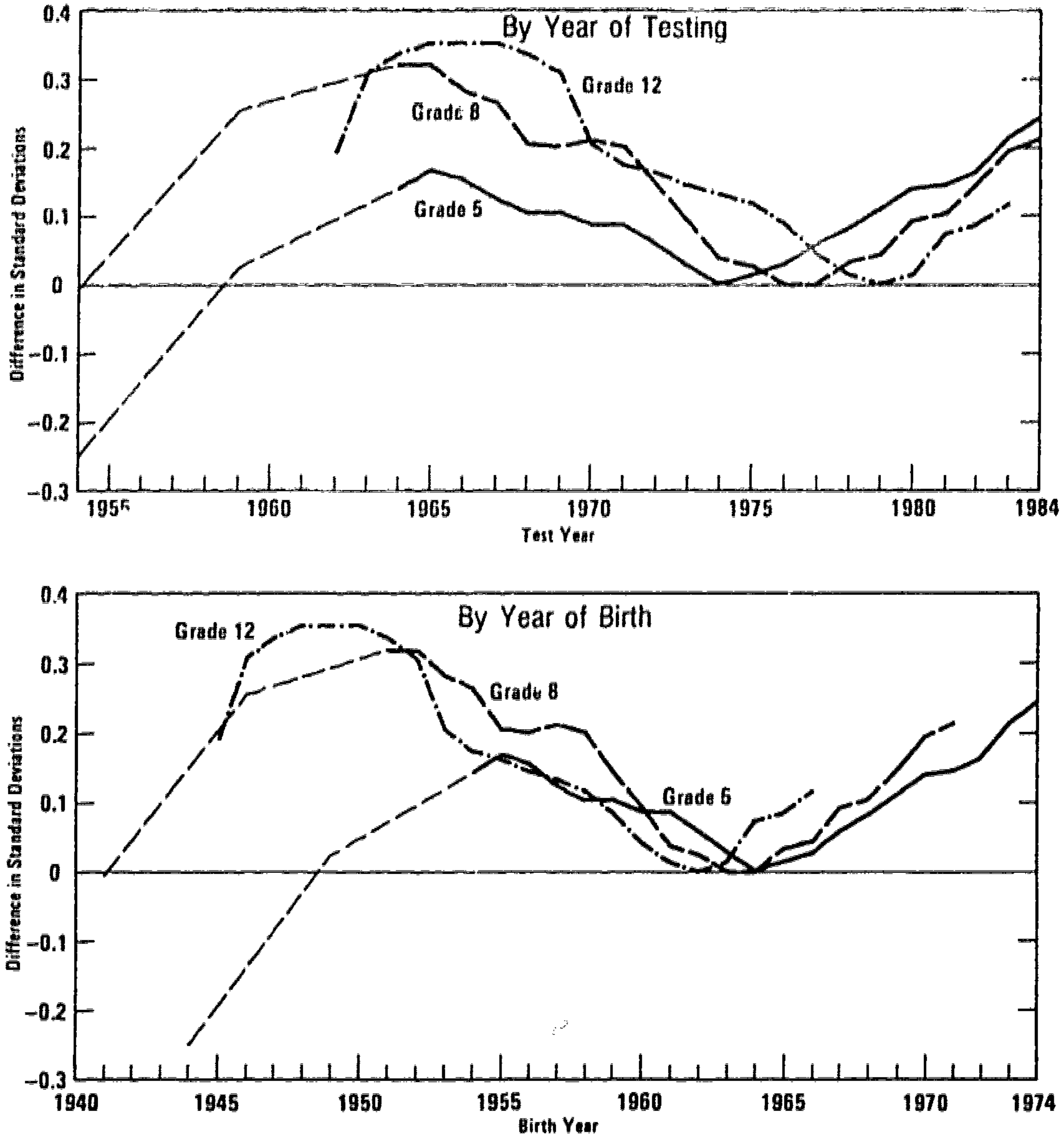
The overall decline in test scores generally ended with the cohorts of children born around 1962 and 1963--that is, with the cohorts that entered school in the late 1960s. Thus, the decline's end first appeared in tests administered in the upper elementary grades in the mid-1970s. Thereafter, it moved into the higher grades at a rate of roughly a grade per year as those birth cohorts aged, reaching the senior high school grades in the late 1970s (see Summary Figure 1). This pattern, however, has gained relatively little attention. Perhaps because of the greater notice accorded to tests at the senior high school level, there has been a widespread misconception that the decline ended only within the past few years.

In fact, subsequent cohorts of children--those entering school in the late 1960s and throughout the 1970s--produced a sharp rise in scores on most, but not all, tests. In the majority of instances in which scores increased, the rise has been steady--with each cohort tending to outscore the preceding one--and often roughly as fast as the decline. As a result, achievement in the elementary grades is now by some measures at its highest level in three decades. At the other extreme, scores on tests administered to high school students, such as the Scholastic Aptitude Test

[2] The average decline on these tests was roughly 0.3 standard deviation.

Summary Figure 1.

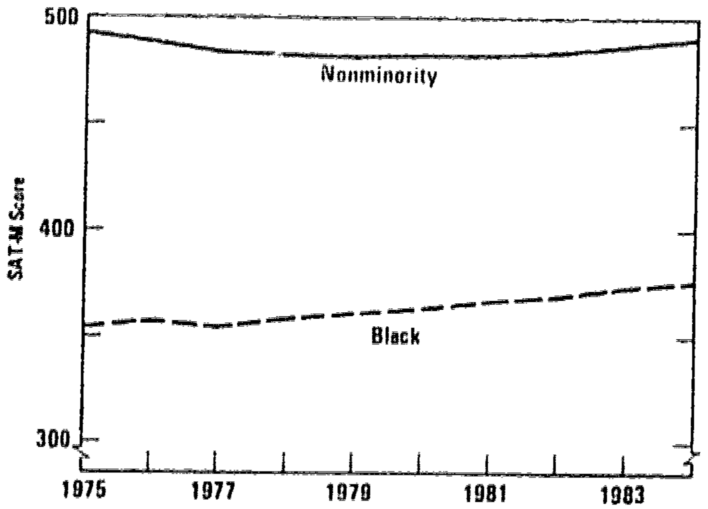
Iowa Average Test Scores, Grades 5, 8, and 12, Differences from Post-1964 Low Point



SOURCES: CBO calculations based on "Iowa Basic Skills Testing Programs, Achievement Trends in Iowa: 1955-1985" (Iowa Testing Programs, unpublished and undated material); A. N. Hieronymus, E. F. Lindquist, and H. D. Hoover, *Iowa Tests of Basic Skills: Manual For School Administrators* (Chicago: Riverside, 1982); "Mean ITED Test Scores by Grade and Subtest for the State of Iowa, 1962 to Present" (Iowa Testing Programs, unpublished and undated tabulations); and Robert Forsyth, Iowa Testing Programs, personal communication, August 1984.

Summary Figure 2.

SAT-Mathematics Scores by Ethnicity: Black and Nonminority Students



SOURCE: The College Entrance Examination Board, "College Board Data Show Class of '85 Doing Better on SAT, Other Measures of Educational Attainment" (press release, The College Board, September 1985).

(SAT), still remain relatively close to their low points of the late 1970s, probably because of the shorter interval since scores began to rise again in those age groups. While it appears that these improvements are occurring at many skill levels, the data raise disturbing questions of whether the improvements on some tests have been larger in the more basic skills than in areas requiring deeper understanding.

Another important issue in the debate over educational achievement is the performance of minority students on standardized tests. Over the past 10 to 15 years--a period that encompassed both declining and improving test scores--the average scores of some minority students rose compared with those of nonminority students. The relative gains of black students appear on every test for which separate trend data for black students are available. Although the gap in average scores between black and nonminority students remains large, it has narrowed appreciably (see Summary Figure 2). [3] Some

[3] On the SAT, for example, the rate at which the scores of black and nonminority scores have converged over the past nine years is comparable to that of the total decline in scores among all students taking the test--a trend that few observers have labeled insignificant.

test results suggest that the scores of black students showed lesser decreases than did those of nonminority students during the final years of the achievement decline, stopped declining earlier, and showed greater improvement during the first years of the overall upturn in scores.

In addition, Hispanic students, who also typically have average scores well below those of nonminority students, showed relative gains over the past decade. The improvement appears to have been greater among Mexican-American students than among other Hispanics. These patterns are less clear-cut, however, because of more limited data, ambiguities in the classification of diverse Hispanic students, and the relatively small number of Hispanics in the test data.

The period since 1970 also included relative improvement of average test scores in certain characteristically low-achieving types of schools and communities. Between 1977 and 1981, mathematics scores on the National Assessment of Educational Progress rose much more sharply in high-minority schools (those with minority enrollments of 40 percent or more) than in other schools. This upturn suggests that the gains of minority students cannot be attributed entirely to those attending schools with low concentrations of minority students. Students in disadvantaged urban schools also showed relative gains in the National Assessments of mathematics and reading. In mathematics, for example, average scores of 9- and 13-year-old students in disadvantaged urban communities rose markedly after 1972, while those of students in other localities rose little or not at all. These relative gains were sizable; by 1981, a fourth to a third of the gap in test scores between disadvantaged urban communities and the rest of the nation had been overcome.

CHAPTER I

INTRODUCTION

Concern about the educational achievement of American students has recently reached its most serious level since the Sputnik-inspired reform era of the 1950s and 1960s. One source of this concern has been a growing public awareness that achievement levels had, by many measures, dropped considerably during the 1960s and 1970s, and that American students compare poorly on achievement tests with their peers in many other nations.[1] A number of prominent reports--such as *A Nation at Risk*--have amplified public concerns about the achievement of American students and called for major changes in the educational system.[2]

The current widespread focus on the educational achievement of students is a part of a much broader concern about the state of American public education. For example, recent reports have cited such issues as apparent declines in the academic qualifications of newly trained teachers; growing shortages of teachers, particularly in certain subject areas; a perceived failure of educational institutions to keep pace with the demands of a technologically changing society; major changes in the characteristics of the school-age population (such as the growing proportion comprising ethnic minorities and children from single-parent families); poor school discipline; and student abuse of alcohol and other drugs.

As concern about the state of public education has grown, Americans have increasingly come to judge the quality of their schools by the results of achievement tests. This trend is apparent from the local to the national

[1] These facts were documented during the 1960s and 1970s, but gained relatively little public attention until the past few years. See, for example, Annegret Harnischfeger and David E. Wiley, *Achievement Test Score Decline: Do We Need to Worry?* (Chicago: ML-GROUP for Policy Studies in Education, 1975); Advisory Panel on the Scholastic Aptitude Test Score Decline, *On Further Examination* (New York: College Entrance Examination Board, 1977); Torsten Husen, ed., *International Study of Achievement in Mathematics: A Comparison of Twelve Countries* (Stockholm and New York: Almqvist & Wiksell and John Wiley & Sons, 1967); and G. F. Peaker, *An Empirical Study of Education in Twenty-One Countries: A Technical Report* (New York: John Wiley and Sons, 1975).

[2] National Commission on Excellence in Education, *A Nation at Risk* (Washington, D.C.: U. S. Government Printing Office, 1983).

level. In some localities, for example, newspapers routinely publish comparisons of the average test scores obtained by students in various schools. On the national level, thi: tendency has taken several forms, perhaps the most salient of which i- the now annual publication by the U.S. Department of Education of the average scores on college admissions tests attained by students in each of the states. Indeed, test scores have come to be used as a national report card on the schools.

Despite the current emphasis on educational achievement, surprisingly little attention has been given to some of the more positive recent trends in the achievement of elementary and secondary school students. The declines of the 1960s and 1970s ended some time ago (as much as a decade ago in the early grades) and have since been superseded by a sizable upturn in test scores. This change has only recently begun to gain widespread recognition and as yet has had little apparent impact on educational initiatives. Similarly, although the large gap in average test scores between nonminority and minority students has been widely acknowledged, the fact that this gap has been slowly but appreciably narrowing in recent years has gained far less attention.

The current heavy reliance on achievement tests makes it critical to gauge recent trends in test scores, to understand the strengths and limitations of test scores as indicators of educational achievement, and to explore their implications for educational policy. This paper assesses recent trends in the achievement test scores of American elementary and secondary school students. It assesses both aggregate trends and variations among groups of students, types of communities, and types of tests. It considers a wide variety of tests in order to ascertain the consistencies underlying the sizable and often unexplained variation in their results. The analysis sl ows that some patterns are reasonably consistent among tests and ther.ore warrant confidence, while others are restricted to one or a few tests and thus should be considered questionable. A forthcoming companion paper, *Educational Achievement: Explanations and Implications of Recent Trends*, evaluates common explanations of the achievement trends and explores the implications of the trends and of their explanations for educational policy.

THE CONTEXT OF THE CURRENT CONTROVERSY

Although states and localities have primary responsibility for public elementary and secondary education--and together provide over 90 percent of the money spent for this purpose by all levels of government--education is a truly national concern. Debate about educational policy thus often emphasizes questions of national interest. For example, although there is surprisingly little evidence about the specific skills and abilities that contribute to success in different occupations, the impact of education on the productivity of the nation's workforce has been an important point of debate at least since the turn of the century.[3] Similarly, the implications of educational policy for national security have often been the focus of attention. Congressional and administration concerns about educational achievement accordingly have often been more far reaching than the relatively small federal role in elementary and secondary education might suggest.

The current national debate about elementary and secondary education--and the participation of the Congress and the administration in the controversy--have numerous historical parallels. For example, current concern that the most able students be given sufficiently challenging curricula has parallels in the 1893 report of the "Committee of Ten"--considered by some historians to be the first major national report on the high school.[4] Similarly, contemporary concern that other students be adequately prepared for the demands they will face after leaving school has precursors in another early national report--*The Cardinal Principles of Secondary Education*, published in 1918--as well as in Congressional and administration actions around tne time of the First World War.[5]

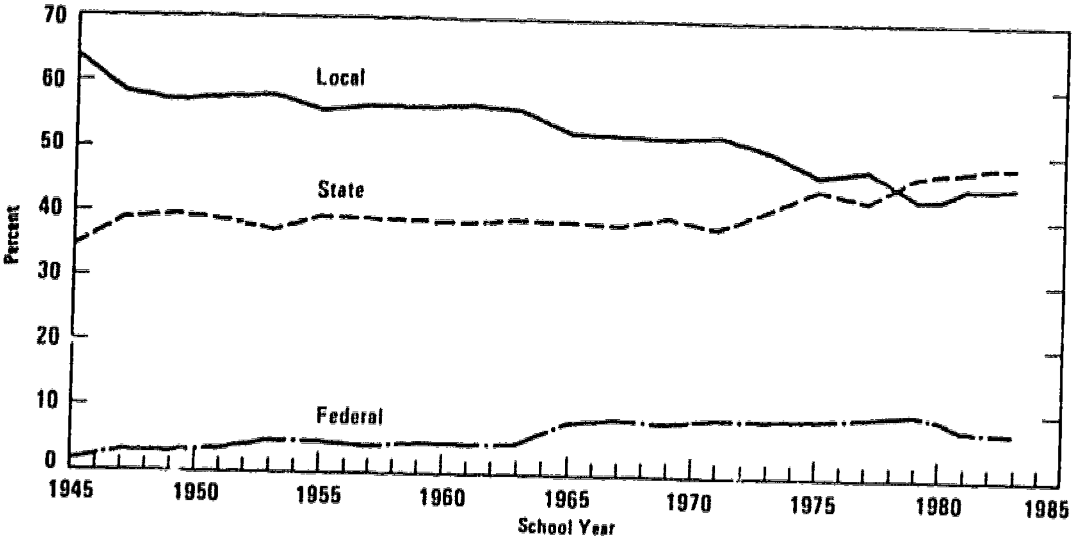
The current wave of concern about educational achievement also mirrors its predecessors in having sparked policy initiatives at all levels of government. The impact of achievement tests, however, in contrast to less specific notions of achievement, has grown much more substantial. Certain uses of tests--for example, minimum-competency tests and other state-

[3] For a description of the technical and economic emphasis in educational debate and programs around the turn of the century, see, for example, David K. Cohen and Barbara Neufeld, "The Failure of High Schools and the Progress of Education," *Daedalus* (Summer 1981), vol. 110, pp. 69-81; and Thomas James and David Tyack, "Learning from Past Efforts to Reform the High School," *Phi Delta Kappan* (February 1983), vol. 64, pp. 400-406. The relevance of such considerations to federal education policies since 1917 is discussed briefly below.

[4] James and Tyack, "Learning from Past Efforts."

[5] *Ibid;* Carl F. Kaestle and Marshall S. Smith, "The Federal Role in Elementary and Secondary Education, 1940-1980," *Harvard Educational Review*, vol. 54 (4) (November 1982), pp. 384-408.

Figure I-1.
Shares of Elementary/Secondary Education Funding by Level of Government



SOURCE: National Center for Education Statistics, *Digest of Education Statistics, 1983-1984* (Washington, D.C.: NCES, 1983), Table 62, and unpublished tabulations.

mandated tests--have grown markedly since the 1970s. Test results now have effects that greatly exceed their impact in earlier eras. These consequences are diverse, ranging from the level of individual students to that of national policy. They include, for example, decisions about the promotion or graduation of individual students; changes in curricula and instruction; the distribution of funds among schools; and changes in educational policy at both the federal and state levels.

Trends in the Federal, State, and Local Roles in Elementary and Secondary Education

Funding for and control over elementary and secondary education was initially a largely local concern. A significant state role began to emerge in the nineteenth century, however, and has continued to grow since.[6] At the

[6] Kaestle and Smith, "The Federal Role."

end of World War II, the states on average supplied about a third of the revenue receipts of public elementary and secondary schools, while local sources provided most of the remainder (see Figure I-1). The state share continued to increase, although erratically, in the post-war years, and has roughly equaled the local share for nearly a decade. 7/ The state share, however, varies greatly; in 1982, it ranged from 9 percent in New Hampshire to 75 percent in Washington and New Mexico and 78 percent in Alaska. 8/

The delineation of state and local responsibilities has also changed over time and varies from one state to another. But both states and localities have clear reasons to be concerned with achievement trends, since they share responsibility for broad questions of curriculum, course requirements, and testing. 9/

The federal role in elementary and secondary education has always been more limited than that of states and localities. Until the end of World War II, the federal government contributed less than 1.5 percent of public school revenues (see Figure 1-1). The federal share climbed to roughly 4 percent over the next decade and remained at that level until the mid-1960s, when it jumped to a range of 8 percent to 9 percent. It remained at that level for about a decade more. From 1977 through 1980, the federal share briefly grew to over 9 percent; thereafter it dropped. By the most common accounting, the federal contribution in the 1983 school year was about $8.7 billion--just under 7 percent of the $126 billion in total public school revenues.

7. That state and local contributions are currently roughly equal is not a matter of controversy, but the precise federal, state, and local shares shown in Figure I-1 are open to question. These estimates, which are from the National Center for Education Statistics, are used because they are perhaps the most common and because they are available for a relatively long historical period; but their use does not represent a judgment about the relative validity of the alternatives. The Census Bureau's Annual Survey of Government Finances yields roughly similar estimates of federal and state contributions but a larger estimate of local funding; the state share is estimated to be a bit lower than the local. Recent alternative estimates from the National Center for Education Statistics show a substantially larger federal share. They do not address the split between local and state sources, however, and are available only for recent years. See National Center for Education Statistics, *Digest of Education Statistics, 1983-84* (Washington, D.C.: NCES, 1983), Table 62; Bureau of the Census, *Finances of Public School Systems in 1983-84*, GF84-No. 10 (Washington, D.C.: U.S. Department of Commerce, 1985), Table B; and National Center for Education Statistics, *Federal Support for Education, Fiscal Years 1980 to 1984* (Washington, D.C.: NCES, 1985).

8. National Center for Education Statistics, *The Condition of Education, 1985 Edition* (Washington, D.C.: NCES, 1985), Table 1.10. Hawaii and the District of Columbia, both of which comprise only a single school district, are excluded from this comparison.

9. See, for example, "Changing Course: A 50-State Survey of Reform Measures," *Education Week*, vol. 4, number 20 (February 6, 1985), pp. 11-30.

The growth in federal funding in part reflected qualitative changes in the nature of federal involvement. Until the 1950s, federal education funds were devoted to a few very narrow purposes. In 1950, for example, federal funds supported only three educational programs, two of which focused on small portions of the school-age population--namely, fiscal assistance to localities affected by federal installations and the education of native American children. Support for vocational education was the sole educational program aimed at a broad segment of students. Moreover, in that year, over half of federal aid was provided, not for educational programs of any sort, but rather for school lunches. 10/ Since then, a variety of laws have greatly broadened federal involvement in elementary and secondary education.

Despite the relatively recent expansion of federal involvement in elementary and secondary education, however, federal efforts to improve the performance of American students date back to the early part of this century. Moreover, the rationale for that involvement has often reflected a common theme: a national interest in the competence and productivity of the labor force produced by the schools.

The Smith-Hughes Act of 1917, which established federal support for vocational education, is often described as the first categorical federal program in elementary and secondary education. One of the aims of this bill, which remains funded to this day, was to improve the skills and productivity of the workforce as a response to international rivalry. 11/ The National Defense Education Act of 1958 (NDEA), which authorized a variety of activities designed to improve instruction in mathematics, science, and foreign languages, had a similar rationale. 12/ Some historians argue that the NDEA had its roots in dissatisfactions with the educational system dating back to the early 1950s. But the launching of Sputnik in 1957 and heightened concern about America's international stature and competitiveness clearly added to the NDEA's momentum and shaped debate about the act. 13/ Some of the concerns of the Smith-Hughes Act were thus mirrored in the NDEA's statement of purpose:

10. Hollis P. Allen, *The Federal Government and Education: The Original and Complete Study of Education for the Hoover Commission Task Force on Public Welfare* (New York: McGraw-Hill, 1950); cited in Kaestle and Smith, "The Federal Role."

11. Kaestle and Smith, "The Federal Role," pp. 388 and 391.

12. Public Law 85-864; 72 Stat. 1580.

13. Kaestle and Smith, "The Federal Role," p. 392.

> The Congress hereby finds and declares that the security of the Nation requires the fullest development of the mental resources and technical skills of its young men and women. The present emergency demands that additional and more adequate educational opportunities be made available... 14/

The large jump in federal funding for elementary and secondary education in the mid-1960s reflected the passage in 1965 of the Elementary and Secondary Education Act (ESEA; Public Law 89-10). ESEA created a broad array of federal education programs, including the compensatory education program that remains the largest single source of federal funds for public schools. 15/ The statement of purpose of the ESEA noted concerns similar to those that motivated Smith-Hughes and the NDEA. Title I accounted for most of the authorized funds, and the act's statement of purpose accordingly focused on an intent to improve the educational opportunities open to disadvantaged students. Nonetheless, the statement also cited concerns more similar to those of Smith-Hughes and the NDEA--the nation's well-being and security. 16/

Similar concerns have been voiced again during the past few years. The report of the National Commission on Excellence in Education, *A Nation at Risk*, asserted that "Our once unchallenged preeminence in commerce, science, and technological innovation is being overtaken by competitors throughout the world. This report is concerned with only one of the many causes...of the problem, but it is the one that undergirds American prosperity, security, and civility." 17/ Another prominent critique of the educational system, produced by the "Task Force on Education for Economic Growth," began by maintaining that improving education is one of the few national efforts that "can legitimately be called- crucial to our national survival." 18/ The Committee Report for the Education for Economic Security Act of 1984, which established a new federal effort to improve

14. Public Law 85-864, Section 101.

15. Title I of ESEA, now Chapter 1 of the Education Consolidation and Improvement Act of 1981.

16. *Elementary and Secondary Education Act of 1965*, H. Rept. 143, 89:1 (1965).

17. National Commission on Excellence in Education, *A Nation at Risk*, p. 5.

18. Task Force on Education for Economic Growth, *Action for Excellence* (Denver: Education Commission of the States, 1983), p. 3.

instruction in mathematics and science, sounded similar themes of national prosperity and security. 19/

In addition to these intermittent direct efforts to improve student performance, the federal government has also taken on an indirect role in this effort by generating, collecting, and disseminating educational information and statistics. Although this role has grown substantially in recent decades, it extends back for more than a century, and it has generally been less controversial than the more direct efforts. The U. S. Department of Education was established in 1867 primarily to gather statistics about education, and that role has continued without interruption to the present. 20/ A National Advisory Committee on Education was established in 1954 to advise the Secretary of Health, Education, and Welfare on educational studies of national concern, and the National Institute of Education was created by the Education Amendments of 1972 (Public Law 92-318). Other major federal efforts to generate, collect, or disseminate information on education accompanied the more direct activities.

Although these information-related activities receive only a small proportion of federal funding for elementary and secondary education, the federal contribution provides a great deal--in some cases, the lion's share--of resources available for carrying them out. In a number of instances, the data generated by the federal government have been unique. For example, all of the truly nationally representative indicators of educational achievement used in this paper--the National Assessment of Educational Progress, the High School and Beyond study, the National Longitudinal Study of the High School Seniors Class of 1972, and Project TALENT--were funded by the federal government.

Recent Policy Initiatives

Numerous recent federal, state, and local efforts to improve educational achievement have reflected these historical patterns. Many state and local governments have made sweeping changes in curricula, high school graduation requirements, testing programs, policies for the certification and

[19] Education for Economic Security Act, S. Rept. 98-151, 98: 2 (1984), p. 1.

[20] The Department of Education was renamed the Office of Education shortly after its establishment and retained that designation until 1979.

compensation of teachers, and other educational policies. 21/ The Administration has emphasized its information-dissemination role in attempts to prompt reforms. 22/ Some of the legislation considered by the Congress (such as the Economic Security Act of 1984) has followed in the tradition of Smith-Hughes and the NDEA in focusing efforts on specific subjects that were considered by the act's proponents to be of particular importance to the nation's competitiveness and security. Other legislation, such as the Secondary Schools Basic Skills Acts, would follow in the mold of Title I of ESEA in funding additional basic-skills instruction for educationally disadvantaged students. 23/

Trends in educational achievement--particularly, the decline of the 1960s and 1970s--have often been cited as a rationale for recent educational initiatives, and some proposals appear to be predicated on assumptions about the causes of those trends. Many of the recent initiatives, however, are not fully consistent with either the trends or the limited information on their causes. For example, some of the proposals do not take into account the nearly uninterrupted increase in test scores in the earliest grades. Others aim primarily at specific curriculum areas--such as the most basic skills--that have shown relatively favorable trends.

Congruence with recent achievement trends is of course only one of many bases on which to ground educational initiatives. Changing a given educational practice, for example, might improve average levels of achievement even if--contrary to common view--that practice did not actually contribute to the decline. But as long as achievement trends are offered as rationales for educational policy changes, the consistency between the proposals and the trends is important to evaluate. Moreover, a more comprehensive view of the trends and their causes allows one to design initiatives to counter the severest problems, to capitalize on recent positive trends, and perhaps to target some of the root causes of both.

21. For example, "Changing Course: A 50-State Survey;" Staff of the National Commission on Excellence in Education, *Meeting the Challenge: Recent Efforts to Improve Education Across the Nation* (Washington, D.C.: Department of Education, November 1983).

22. For example, National Commission on Excellence in Education, *A Nation at Risk;* U.S. Department of Education, *State Education Statistics: State Performance Outcomes, Resource Inputs, and Population Characteristics, 1982 and 1984* (January 1985); U.S. Department of Education, *Indicators of Education Status and Trends* (January 1985).

23. S. 508, introduced by Senator Bradley, and H.R. 901, introduced by Representative Williams.

CHAPTER II

UNDERSTANDING MEASURES OF EDUCATIONAL ACHIEVEMENT

In recent years, the use of standardized tests as indicators of achievement has been burgeoning. These tests are diverse, including minimum-competency tests (MCTs), college admissions tests, and "norm-referenced" achievement tests. All of them, however, have one common characteristic: they apply a uniform measure to gauge the performance of diverse students in a wide variety of settings.

Many advantages of standardized tests over alternative measures--such as grade-point averages and locally developed tests--are obvious. On the other hand, ·/hile the limitations of standardized tests are less obvious, they can be severe. 1/

Perhaps the most important strength of standardized tests is that they can be freed of much of the subjectivity that can plague suc.h alternative measures as teachers' grades or class rank. They can also avoid other extraneous variations in evaluations of student performance, such as differences in grading standards. If appropriately designed and scored, standardized tests can be made comparable over time and can yield useful information about trends that is unavailable from other sources. Standardized tests can also be designed to provide valid indices of specific aspects of achievement. They can be designed, for example, to differentiate among particularly high- or low-achieving students, tap specific types or levels of skills, or provide comparable information on the performance of students in different grade levels.

Despite these strengths, the seemingly straightforward information provided by standardized tests often masks considerable complexity and

1. Although many of the key issues in testing are technically complex, this chapter provides a largely nontechnical description for readers who are unfamiliar with testing and statistics. Readers desiring a more detailed and technical discussion of the issues discussed in this chapter are referred to "Testing: Concepts, Policy, Practice, and Research," a special edition of *The American Psychologist*, vol. 36, (October 1981), and, in particular, to Bert Green, "A Primer of Testing," pages 1001-1012 in that volume, on which parts of this chapter draw substantially.

ambiguity. One indication of the limitations of standardized tests is the often marked disparities in the results they yield (see Chapter III). This divergence can reflect differences in the purposes and construction of the tests, such as discrepancies in content, level of difficulty, or test format. On the other hand, its causes are often poorly understood, and it can also appear when tests are apparently similar.

The limitations of standardized tests are particularly severe when they are used to compare schools, districts, states, or other aggregates--as they increasingly have been in recent years. Such comparisons are difficult and can be seriously misleading. Standardized measures in themselves can remove only some, but not all, of the extraneous variation among groups. For example, comparisons among jurisdictions can be seriously biased by differences in dropout rates, the composition of the school-age population, rules governing exclusion of certain groups from testing, and the closeness of the match between the test and curricula.

Using standardized tests to gauge trends is also especially problematic. To assess trends accurately, test results must be made comparable from one testing to the next. This process is more difficult than it might seem (as is described below). When test results are not made fully comparable, estimates of trends can be seriously distorted.

EDUCATIONAL TESTS VERSUS EDUCATIONAL ACHIEVEMENT

Although popular accounts often treat test scores as synonymous with educational achievement, the two are in fact very different. In most cases, tests are not direct and comprehensive measures of educational achievement. Rather, they are proxies, or substitutes, for such ideal but generally unobtainable measures, varying markedly in how much they differ from the ideal. The choices made in designing that substitution are many and have a large impact on the results obtained.

Perhaps the best way of understanding an educational test is to consider it an activity, the performance of which is intended to predict some other performance or attribute that is more difficult to measure directly.[2] In some instances, what the test predicts cannot be directly

[2] Douglas Coulson of the Office of Technology Assessment suggested this metaphor.

measured because it lies in the future (such as performance in subsequent schooling or work). In other cases, the test is a proxy for a present characteristic of the student--such as mathematics achievement--that is difficult or impossible to measure completely.

An example of a test that differs markedly from the activities for which it is a proxy is the Scholastic Aptitude Test (SAT). The SAT is intended to predict students' performance in college, and much of the work gauging that test's value assesses the correlations between SAT scores and freshman-year college grades.[3] Taking the SAT, however, is an activity very different from most of those in which college students must succeed. Those students who do well on a multiple-choice examination are not necessarily those who can concentrate through an hour-long lecture, discipline themselves to do considerable amounts of reading over a long period of time, or write well-organized and fluent term papers. For this reason, the SAT predicts college performance only imperfectly.

While most achievement tests, unlike the SAT, are intended to assess the present knowledge or other current attributes of students rather than their future performance, striking differences can still exist between the activities constituting the test and the real-life skills for which they are proxies. For example, many tests use a multiple-choice format, in part because of ease of scoring. The corresponding tasks in real life, however often involve quite different skills--writing prose, solving a mathematics problem without any clue about possible solutions (and even without a clear statement of the problem), inferring or hypothesizing explanations of events, assessing the logic and persuasiveness of arguments, and so on.

Given these differences between tests and the corresponding real-life activities, creating a test--and understanding the results of one already administered--raise several sets of questions:

- What is the test's purpose, and what real-life skills are of interest?

- What test activities--at what level of skill and in what format--will be used to represent those real-life skills?

[3] Hunter M. Breland, *Population Validity and College Entrance Measures*, Research Monograph Number 8 (New York: The College Board, 1979).

- To what extent is performance on the test actually a reasonable gauge of the real-life skills of interest? and

- How are the scores scaled and reported?

IMPORTANT CHARACTERISTICS OF EDUCATIONAL TESTS

Many characteristics of educational tests have a major impact on the results those tests yield. This section describes some of the most important test characteristics and illustrates their impact on test results.

What Is the Purpose of the Test?

Most of the commonly discussed educational tests are designed to achieve one of three purposes:

- Ascertain whether students have acquired specific skills or information;

- Rank students in terms of their knowledge or skills; or

- Predict subsequent performance. [4]

Tests That Ascertain Whether Students Have Acquired Specific Skills or Information. Among the tests intended to gauge whether students have acquired specific skills or knowledge are the *minimum-competency tests* (MCTs) now used by many states and localities as criteria for promotion, graduation, or remedial services. The content of these tests generally reflects a judgment about the skills and knowledge that most or all students should master, and thus the level of difficulty is often deliberately quite low. Because tests of this type entail comparing a student's performance with a concrete criterion for achievement, they are called *criterion-referenced tests.*

[4] Although using test results to compare or rank jurisdictions--schools, districts, and states--is currently enjoying a vogue, none of the tests reported in this paper was designed for that purpose. The difficulties that arise in using them to that end are discussed later in this chapter.

How items are typically selected for inclusion in criterion-referenced tests has important implications for comparisons among groups of students and for the assessment of achievement trends. Whether an item is selected depends primarily on the extent to which it represents an aspect of the criterion or skills to be taught. For that reason, assuming that the item has no other problems (such as ambiguous wording), the proportion of students correctly answering it can be irrelevant. In the case of MCTs, one might find both test items that most students answer correctly and a large number of very high scores. These results would reflect the typically low level of achievement (the "minimum competency") used as a criterion and would simply be interpreted as evidence that the schools are successfully imparting that particular set of skills. 5/

When criterion-referenced tests such as MCTs include many questions that most students answer correctly (or incorrectly), comparisons between high- and low-achieving students often become very difficult to interpret. For example, if the test is relatively easy, high-scoring students will score near or at the maximum. Even so, some of their scores will be lower than they might otherwise be, since the absence of more difficult items on the test leaves no way for the higher-achieving students to distinguish themselves from others. This is often referred to as a *ceiling effect*; the opposite is called a floor effect.

One result of the ceiling effect in some MCTs is that when scores are generally increasing--as has been the case with many tests in recent years--they will tend to show low-achieving groups as gaining on higher-achieving groups, even when all groups are actually improving comparably. Because of the ceiling, the scores of the higher-achieving groups cannot increase proportionately to mirror their true improvement.

Tests That Rank Students in Terms of Their Knowledge or Skills. In contrast to MCTs, those achievement tests that for years were the standard in elementary and secondary schools rate students by comparison to the performance of other students, rather than by comparison to an absolute achievement criterion. For example, a student's performance might be reported as being at the 75th percentile, meaning that it exceeded the achievement of three-fourths of all students.

5. A very high success rate on an MCT, however, may be taken as a sign that the test is no longer serving its function, since it no longer indicates skills that need improvement. That is, it might call the achievement criterion itself into question. New Jersey, for example, recently decided that its MCT needed replacement with a more difficult test for this reason. *Statewide Testing System, New Jersey Public Schools* (Trenton: New Jersey State Department of Education, January 1983).

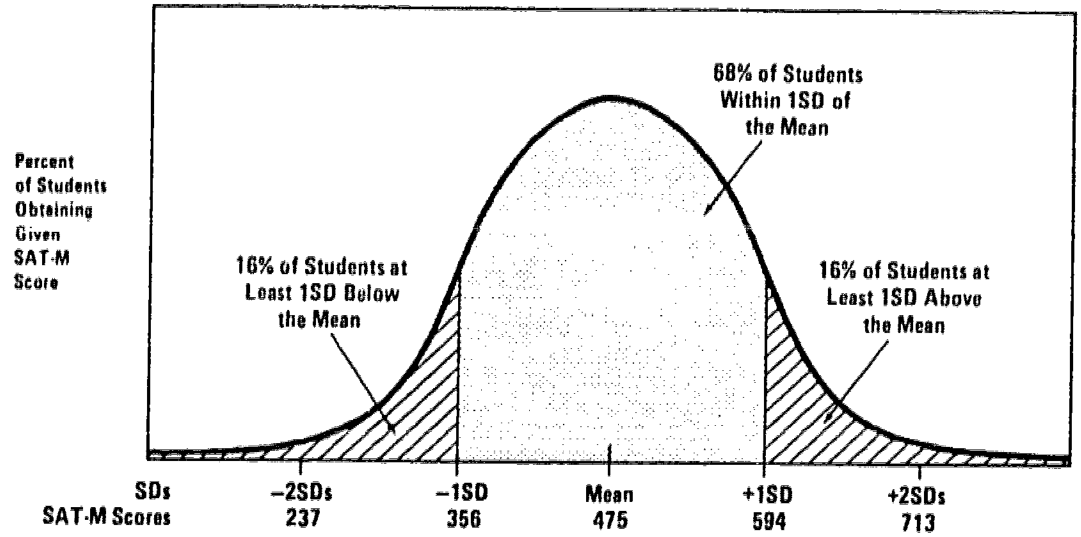
The distribution of scores with which students are compared is called the "norms," and such tests are therefore called *norm-referenced.* The norms are typically derived from a national sample of students and are generally revised infrequently--typically, at intervals of seven years or so. Revision of the norms--often called "renorming"--generally entails both revision of the test itself and retesting with a new national sample. One technique, for example, is to revise the test and then to administer both the old and new versions to a large national sample of students. This approach provides both a new set of norms and a measure of the extent to which changes in scores reflect the revision of the test itself rather than a change in achievement.

Norm-referenced tests are often relatively free of the floor and ceiling effects that can plague interpretation of MCTs. Since norm-referenced tests are designed to rank students, they typically must be easy enough to differentiate among low-achieving students but difficult enough to discriminate at the high end of the achievement distribution.

Performance on norm-referenced tests can be scored in many ways, and one common scale--*standard deviations*, or *SDs*--is especially important in understanding the trends reported in later chapters. The reporting of scores in terms of standard deviations allows the comparison of trends among many different tests. The distribution of scores on norm-referenced tests typically resembles the "normal" or bell-shaped curve--that is, many scores are clustered around the average score, while smaller numbers of students obtain scores farther from the average (see Figure II-1). 6/ When scores are distributed that way, the standard deviation is a convenient measure of how far a given student's score is from the average. A student scoring 1 standard deviation above the average has exceeded the scores of about 84 percent of all students, and a student with a score 2 SDs above the average has scored above 97.7 percent of all students. (The measure is symmetrical, so that a student scoring 1 SD below the mean has exceeded the scores of about 16 percent--100 minus 84--of all students.)

6. Test scores generally do not entirely conform to the bell-shaped curve, but the departures from the normal curve are often small and relatively unimportant for many purposes. The distribution of SAT scores, for example, typically is a bit flatter near the mean than is the normal curve, as a result of correlations between items on the test. It is also often slightly skewed toward the higher end of the scale, although this varies with the subtest and particular administration of the test. Finally, SAT scores are bounded at both ends, with a minimum of 200 and a maximum of 800. (William Angoff and Gary Marco, Educational Testing Service, personal communication, March 1986).

Figure II-1.
Hypothetical Test Results Expressed in Standard Deviations (SDs), Based on the SAT-Mathematics (SAT-M)



SOURCE: Adapted from the 1984-1985 SAT-M scores, *National College-Bound Seniors, 1985* (New York: The College Board, 1985).

NOTE: The SAT is only approximately normal, although the deviations from normality are relatively minor for most purposes (see the text).

Tests That Predict Future Performance. A variety of tests--including college-admissions tests such as the SAT and the American College Testing Program (or ACT) tests--are designed to predict future performance rather than to assess current levels or past acquisition of skills.

The SAT and ACT outwardly resemble the norm-referenced achievement tests in many respects, and the trends shown by the two types of tests can in some respects be interpreted similarly. Moreover, the distribution of scores is nearly "normal," or bell-shaped, and thus students' scores can be expressed in terms of the number of standard deviations from the average. Accordingly, they largely avoid ceiling and floor effects.

Despite their outward similarity to norm-referenced achievement tests, however, college-admissions tests are not necessarily indicators of achievement. The value of such a test lies in its ability to predict performance in college. A student's current level of achievement is only one of many attributes that might predict future performance. Alternatives might include, for example, general problem-solving abilities, attention span, or such cognitive measures as fluid intelligence or spatial visualization. Whether a test used to predict college performance relies substantially on current achievement rather than other attributes thus depends on whether one believes--or can demonstrate--that current achievement is a better predictor than are those alternatives. In fact, the SAT is quite dissimilar from most achievement tests. The mathematics portion, for example, is intended to "depend less on formal knowledge than on reasoning" and is deliberately not closely tied to secondary-school mathematics curricula. The College Board has repeatedly protested the misuse of the SAT as a measure of the effectiveness of elementary and secondary education.[7] The ACT, on the other hand, in many respects resembles achievement tests more closely than does the SAT and is intentionally more closely tied to secondary-school curricula.[8]

What Skills and Skill Levels Will Be Assessed?

Once the purpose of a test is decided, decisions must be made about the actual test content--the specific skills and knowledge to be assessed and the le el of difficulty to be targeted. Unless the purpose of a test is extremely narrow--for example, testing proficiency in two-digit subtraction problems--these decisions are vexing and their solutions ambiguous. For example, many diverse skills are subsumed by broad categories such as "reading" or "mathematics," even at the elementary school level. Test makers must choose among these skills and decide the relative emphasis that each of those chosen should receive.

[7] Advisory Panel on the Scholastic Aptitude Test Score Decline, *On Further Examination* (New York: The College Entrance Examination Board, 1977), pp. 3 and 5; Statement by Daniel B. Taylor, Senior Vice President, The College Board, before the Subcommittee on Elementary, Secondary, and Vocational Education, Committee on Education and Labor, U. S. House of Representatives, January 31, 1984.

[8] Personal communication, Mark Reckase, American College Testing Program, January 1985.

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 x 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, [illegible] with scores on other similar tests. 10/

9. More discussion of [illegible] 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 stud... in the low-scoring group went from having 20 percent to 40 perc-nt 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 gair. 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 adcpting 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.

- College-admissions tests--the Scholastic Aptitude Test (SAT) and the American College Testing Program tests (ACT);
- Most tests in the National Assessment of Educational Progress (NAEP);
- 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;
- Periodic norming data from commercial standardized tests of elementary and secondary achievement;
- 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
- 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).

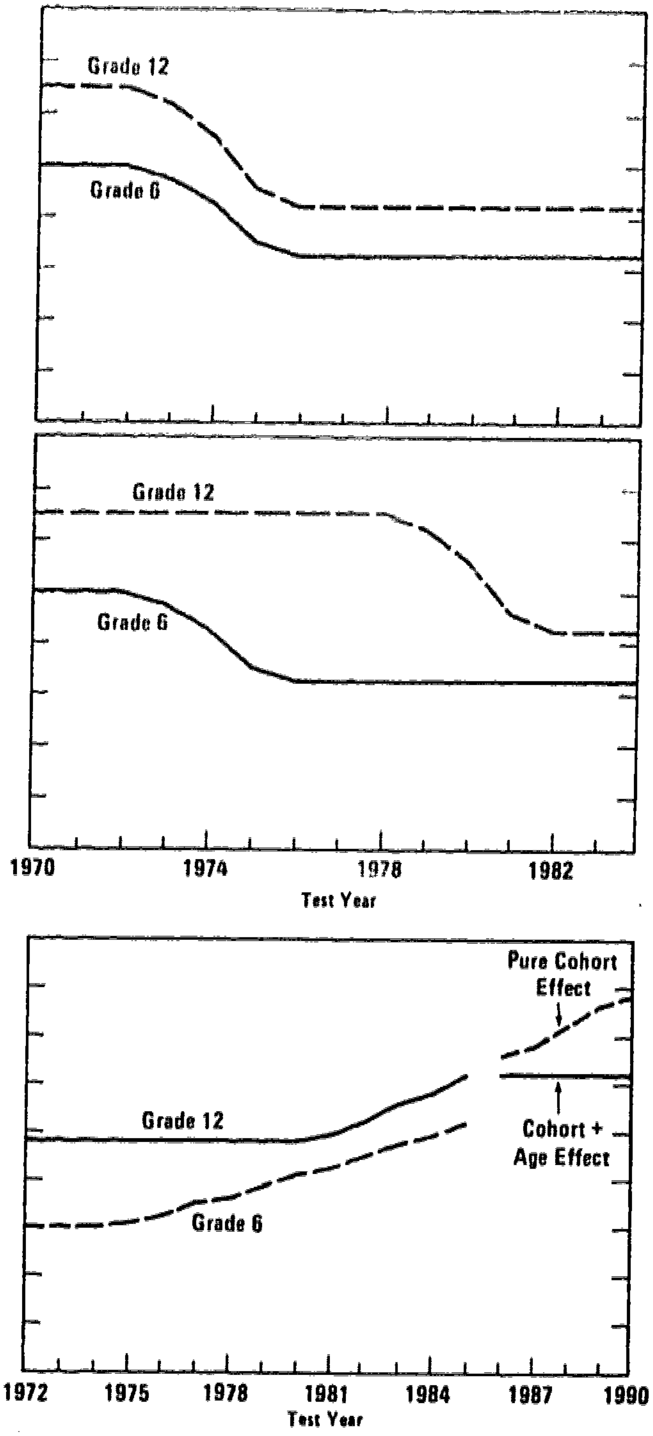


Figure III-1.

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

	Onset		End	
Test	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.

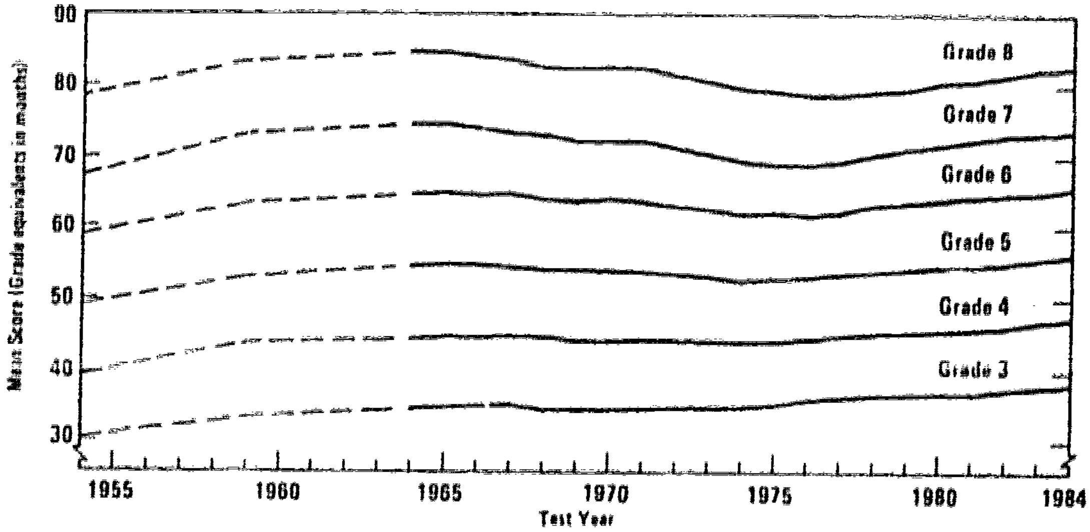
SOURCES: Hunter M. Breland, *The SAT Score Decline: A Summary of Related Research* (New York: The College Board, 1976), Table 1; *National College-Bound Seniors, 1985* (New York: The College Board, 1985); L. A. Munday, *Declining Admissions Test Scores* (Iowa City: American College Testing Program, 1976), 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).

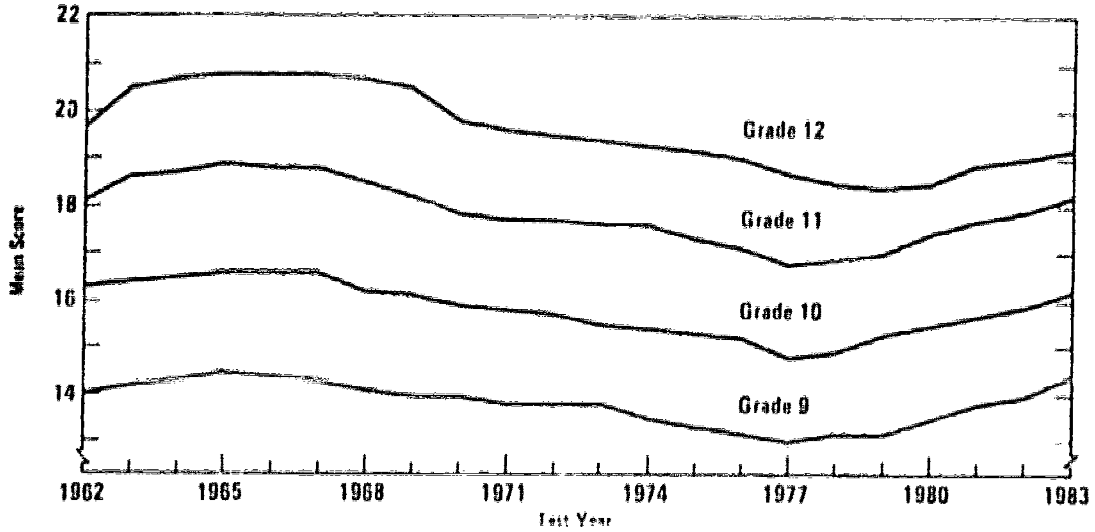
series. 10/ The progression through the grades is somewhat erratic--perhaps because of various unexplained year-to-year fluctuations in average scores--and is therefore not always apparent from a comparison of a few adjacent grades from a single test. The pattern becomes clearer, however, when a range 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.

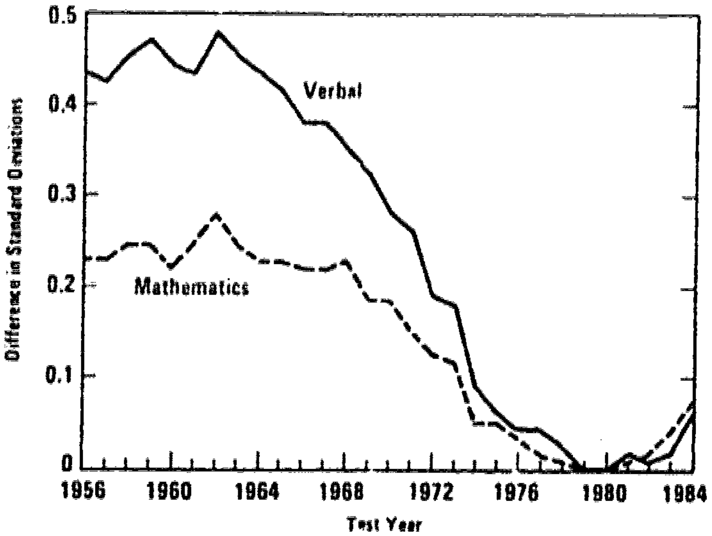
[11] One salient exception to this pattern is the ACT, which reached its low point a few years earlier.

Figure III-3.
ITED Composite Scores, Iowa Only, by Test Year and Grade at Testing



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

Figure III-4.
Average SAT Scores, by 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, 1976), Table 1; and the College Entrance Examination Board, *National College-Bound Seniors, 1985* (New York: The College Board, 1985).

(The extent to which a number of tests conform to this pattern is explained in Appendix B.)

The widespread misconception that the achievement decline ended only within the past few years thus probably stems from the greater attention paid to tests administered to high-school juniors and seniors--in particular, the SAT. The tests that showed an end of the decline taking place a decade or more ago are those given to young children, and they have been the focus of considerably less attention.

How Large Was the Decline?

The severity of the decline in achievement can be illustrated in two ways: by examining the actual level of achievement shown by typical students in each of two years (a criterion-based or absolute standard), or by comparing the achievement of a typical student in one year to the distribution of achievement in some other year (a normative standard). 12/ This section applies both standards.

The Size of the Decline Relative to a Normative Standard. The few test series reporting trend data in normative terms suggest that, at grades 6 and above, the decline averaged about 0.3 standard deviation over the entire period of the decline (see Table III-2). 13/ This average indicates that the median student at the end of the decline would have scored at about the 38th percentile at the beginning of the decline. The severity of the decline varies so greatly, however, that a single average has little value. At one extreme, the largest decline in the measures considered here was 0.55 standard deviation, placing the median student at the end of the decline roughly at the 29th percentile before the decline began. (The two largest declines, however, were on college admissions tests--the SAT and ACT--and were substantially exacerbated by changes in the composition of

[12] Most often, the "typical" score is the mean or median in each year. Since the characteristics of the groups taking most tests change over time, trends in these typical scores in part reflect changes in student characteristics, rather than only changes in the achievement of a student with any given characteristics.

[13] See Chapter 2 for an explanation of standard deviations.

The numbers here do not adjust the SAT trends for "scale drift," a gradual drop in the level of the difficulty of the test that led to an understatement of the SAT decline until the early 1970s. That adjustment was not made because of a lack of information about the severity or direction of any changes in difficulty since that time. If the adjustment is made, however, the conclusions of this section are unaltered; the average decline remains about 0.3 standard deviations, and the maximum decline is in the range of 0.57 to 0.60 standard deviations rather than 0.55.

TABLE III-2. SIZE OF THE ACHIEVEMENT DECLINE, INDICATED BY SELECTED TESTS AT GRADE 6 AND ABOVE
(Including only tests spanning all or nearly all of the decline)

Test	Subject	Total Decline (Standard Deviations)
	Specific Tests	
SAT a/		
Largest	Verbal	0.48
Smallest	Mathematics	0.28
Iowa Grade 12 (ITED)		
Largest	Reading b/	0.40
Smallest	Mathematics	0.27
Iowa Grade 10 (ITED)		
Largest	Reading b/	0.32
Smallest	Natural Science	0.2
Iowa Grade 8 (ITBS)		
Largest	Mathematics	0.47
Smallest	Vocabulary	0.26
Iowa Grade 6 (ITBS)		
Largest	Mathematics	0.38
Smallest	Vocabulary	0.10

(Continued)

TABLE III-2. (Continued)

Test	Subject	Total Decline (Standard Deviations)
ACT		
Largest	Social Studies	0.55
Smallest	Science	-0.06[c]
	All Tests in Table	
Average		0.31
Minimum		-0.06
Maximum		0.55

SOURCES: CBO calculations based on Hunter M. Breland, *The SAT Score Decline; College Board, National College-Bound Seniors, 1978* and *1985;* 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); Robert Forsyth, personal communication, August, 1984; A. N. Hieronymus, E. F. Lindquist, and H. D. Hoover, *Iowa Tests of Basic Skills: Manual for School Administrators* (Chicago: Riverside, 1982); L. A. Munday, *Declining Admissions Test Scores;* and American College Testing Program, *National Trend Data for Students Who Take the ACT Assessment.*

NOTE: Alternate grades (7, 9, 11) omitted for clarity.

a. SAT scores are not adjusted for scale drift. Research indicates that the first part of the decline is understated by perhaps 0.09 standard deviations because of scale drift. The extent and direction of scale drift over the past decade is not yet known, however.

b. This reflects the "Interpretation of Literary Materials" test. Reading skills also are tapped by the other tests in the ITED battery.

c. Negative numbers represent an increase in average scores.

the test-taking group.) 14/ At the other extreme, the ACT natural science test actually showed a trivial increase during the years of the general decline. Thus, a different mix of tests--or a larger and more representative sample of tests--might have yielded a very different average size of the decline.

Some of the variability in the size of the decline stems from known causes, such as the age of the students tested and changes in the composition of the groups of students taking the test. On the other hand, much of the variation appears to stem from unknown factors or from considerations that lie largely outside of the scope of this report, such as decisions about the specific skills and knowledge to be tested.

The Size of the Decline Relative to an Absolute Standard. Although the apparent severity of the decline varies with the absolute achievement criterion chosen, the average decline was clearly large enough by many standards to be educationally significant.

The best criterion-based gauge of the achievement decline is probably the National Assessment of Educational Progress (NAEP). The NAEP reflects representative samples of the national population of students, tests students at the elementary, junior-high, and senior-high levels, and encompasses a wide array of substantive areas and types of skills. Moreover, actual test items from all of the NAEP assessments have been published, along with the percentages of students of different ages answering each item correctly. This information provides an intuitively clear view of students' levels of achievement. 15/

Even the NAEP, however, should be used to illustrate the types of skills that deteriorated rather than to indicate the total magnitude of the decline. Because of the timing of NAEP assessments, most of them understate the severity of the decline, in some instances probably by a very large margin. The NAEP began with a science assessment in 1969, with initial assessments in other subjects starting over the following several

14. Advisory Panel on the Scholastic Aptitude Test Score Decline, *On Further Examination* (New York: The College Entrance Examination Board, 1977); L. A. Munday, *Declining Admissions Test Scores*. The impact of compositional changes is discussed in Congressional Budget Office, *Educational Achievement: Explanations and Implications of Recent Trends* (forthcoming).

15. While annually equated tests provide much clearer information on trends, no such tests have been tabulated in a way that facilitates comparison with an absolute achievement criterion.

years. The most recent published assessments were largely carried out between 1976 and 1981. Therefore, the NAEP trends exclude varying portions of the early part of the decline and probably often mask the later decline by mixing with it upturns in achievement that have occurred in recent years. 16/

Although the NAEP tests students at ages 9, 13, and 17, this section describes the results among 17-year-olds, because the trends among 17-year-olds are likely to include fewer, if any, years of the recent upturn in achievement. 17/ (Comparable information on ages 9 and 13 appears in Table III-3.)

Mathematics. Between 1972 and 1977, the proportion of NAEP mathematics items answered correctly by 17-year-olds dropped from 64.0 to 60.4 percent (see Table III-3). While this decline appears modest, it occurred over a time span that was probably less than half of the total period of decline and also masks more substantial deterioration of performance on certain important types of items. 18/ In addition, the rate of success on certain types of items was remarkably poor in both years. One NAEP computation item, for example, asked: "Express 9/100 as a percent." The proportion of 17-year-olds answering this item correctly dropped eight percentage points over the five years, from 61 percent to 53 percent. Similar results were obtained by a problem that asked: "A hockey team won 5 of the 20 games it played. What percent of the games did it win?" Another problem required students to use a simplified electrical bill to determine the cost per kilowatt if 606 kilowatts produced a bill of $9.09. The proportion of students succeeding on this item fell from 12 percent in 1973 to 5 percent in 1978. 19/

16. Because NAEP assessments are carried out at intervals of four or five years, the ends of the decline in each of them cannot be pinpointed. This precludes estimating any recent increase in each series and disentangling it from the estimates of the preceding downturn.

17. In interpreting the examples given below, it is important to bear in mind that only 17-year-olds still in school were tested in the National Assessment. As a result, the NAEP results are likely to overestimate--perhaps by a sizable margin--the average level of achievement attained by the entire cohort of 17-year-olds.

18. The subsequent interval from 1977 to 1981 showed little change, but it probably brackets the end of the decline and therefore includes some of the subsequent upturn.

19. These and the following mathematics examples are taken from National Assessment of Educational Progress, *Changes in Mathematical Achievement, 1973-1978* (Denver: NAEP/Education Commission of the States, 1979).

TABLE III-3. SUMMARY OF NATIONAL ASSESSMENT RESULTS IN THREE SUBJECTS, AGES 9, 13, AND 17
(Average percent of items correctly answered)

Subject	Age 9	Age 13	Age 17
Mathematics[a]			
1972	56.7	58.6	64.0
1977	55.4[b]	56.6[c]	60.4[c]
1981	56.4	60.5[c]	60.2
Reading[d]			
1970	64.0	60.0	68.9
1974	65.2[c]	59.9	69.0
1979	67.9[c]	60.8	68.2
Science			
1969	61.0	60.2	45.2
1972[e]	59.8[c]	58.5[c]	42.5[c]
1972[f]	52.3	54.5	43.4
1976	52.2	53.8	46.5[c]

SOURCES: CBO calculation based on National Assessment of Educational Progress, *Three National Assessments of Reading* (1981), Tables 2, 4, and 6. Mathematics: *The Third National Mathematics Assessment: Results, Trends, and Issues* (1983), Tables 5.1 and 5.2, and *Mathematical Technical Report: Summary Volume* (1980), Tables 2, 3, and 4; and *Three National Assessments of Science* (1978), Table A-1 (Denver: NAEP/Education Commission of the States).

[a] 1977 and 1981 scores reflect all items used in those two assessments. 1972 scores are obtained by subtracting from 1977 scores the change between 1972 and 1977 on all items used in those two years.

[b] Change from preceding test marginally significant, p less than .10.

[c] Change from preceding test statistically significant, p less than .05.

[d] All scores reflect all items used in all three years.

[e] Reflects only test items shared with 1969.

[f] Reflects only test items shared with 1976.

Student achievement also dropped on certain NAEP items that were less tied to concrete applications. For example, the proportion of 17-year-old students correctly finding a missing numerator in an equivalent fraction fell from 82 percent to 72 percent. The proportion who could solve for x and y in a system of linear equations dropped by a third, from 18 percent to 12 percent.

On the other hand, success rates on some items did not decline--an optimistic note that is tempered by the fact that in many instances the rate was poor in both years. For example, in both 1973 and 1978, about 20 percent of students successfully graphed the equation $y = 2x + 1$. About 15 percent and 12 percent could identify the slope and intercept, respectively, of the equation $2y = 5x - 8$. Five percent ascertained the equation of a line when both the x- and y-intercepts were given.

Reading. In contrast to mathematics, the first three NAEP reading assessments showed *no* substantial overall decline in the achievement of 17-year-olds (see Table III-3). This pattern is inconsistent with a variety of other tests that showed substantial declines in reading and reading-related skills. The results of those other tests, however, have not been published in a form that permits comparison with a concrete achievement criterion.

On the other hand, a decline was apparent in one of the specific reading skill areas tapped by the NAEP--inferential comprehension (that is, comprehension that requires going beyond the information explicitly stated in the question). This discrepancy is discussed in a later section.

Science. Over the seven-year span covered by the first three NAEP assessments of science--1969, 1972, and 1976--the average score of 17-year-olds dropped 4.6 percentage points, or about 10 percent (see Table III-3).

As in the case of the mathematics assessment, the low success rate on certain items is as striking as the decline. One NAEP item, for example, asked, "Which of the following happens when any combustion reaction takes place?" The correct choice--that heat is evolved--was selected by about 68 percent of 17-year-olds in 1969 and by about 54 percent in 1977. Another item asked for explanation of the statement that the relative humidity is 50 percent. About 47 percent of students in 1969 and 42 percent in 1977 selected the correct answer--"The atmosphere contains half as much water as it could at its present temperature." 20/

Social Studies. The NAEP citizenship and social studies assessments in 1968, 1971, and 1975 showed sizable declines in the proportion of 17-year-olds correctly answering some items assessing knowledge of the Constitution, the structure and function of government, the political process, and international affairs. A smaller number of items, however, showed increases.

In one example, the proportion of students answering that a statement of civil rights can be found in the Constitution dropped from 85 percent to 81 percent between 1971 and 1975. 21/ The proportion correctly answering the question "The Congress of the United States is made up of two parts. One part is the House of Representatives. What is the other part?" fell from 94 percent to 88 percent from 1968 to 1975. (The proportion choosing the most popular incorrect answer--the Supreme Court--doubled to 8 percent during that period.) The proportion recognizing that the Congress was part of the legislative branch of government dropped during the same time, from 84 percent to 74 percent. Fifty-four percent of 17-year-olds in 1968, but only 35 percent in 1975, recognized that the circumstance of a state having more Senators than Representatives occurs as a result of low population. The proportion able to define "democracy" declined from 86 percent to 74 percent between 1968 and 1975.

THE RECENT UPTURN IN ACHIEVEMENT

Since the end of the achievement decline, the general trend has been a marked upturn in average achievement. In some instances, the rate of increase has been comparable to or even greater than the rate of decrease during the later years of the decline, and average scores on some tests have approached or exceeded their predecline high points. Moreover, the pattern

20. National Assessment of Educational Progress, *Three National Assessments of Science: Changes in Achievement, 1969-77* (Denver: NAEP/Education Commission of the States, 1978).

21. This and the following examples are taken from National Assessment of Educational Progress, *Changes in Political Knowledge and Attitudes, 1969-76* (Denver: NAEP/Education Commission of the States, 1978).

of the trends among tests administered at different ages suggests that some of the test batteries that have seen only a modest upturn to date--most notably, the SAT--might show marked increases in the next several years.

In contrast, there are a number of important exceptions to this optimistic picture. Scores on the American College Testing Program college admissions tests have yet to turn up substantially. A statewide assessment program in Pennsylvania has shown stable scores in the lower grades and slight deterioration at the secondary level in recent years.[22] The California statewide assessment also has shown no upturn among seniors, though it has shown increases in the lower grades.[23]

Much of the variation in recent trends appears linked to the age of the students: tests given to older students have generally increased less in total than have those administered to younger children. At one extreme, some tests administered in the elementary grades have risen to their highest levels on record--a span of as much as three decades. At the other pole, the generally better known tests administered in the high school grades (such as the SAT) have generally shown more modest gains.

The smaller total upturn to date in the higher grades appears to reflect the shorter time since the upturn began in those grades, rather than a lesser rate of improvement. The upturn, like the end of the decline, shows a cohort pattern, and fewer of the cohorts producing rising scores have yet reached the higher grades. (The relationships between age and the subsequent upturn are discussed further in Chapter IV.)

This pattern suggests that scores on tests administered in the higher grades might rise further in the coming years. That is, the cohorts responsible for the most recent rise in scores in the lower grades might be expected to produce similar gains as they move through the higher grades. The cohort pattern notwithstanding, however, any number of factors could cause future trends in the higher grades to diverge from the recent trends produced by those same cohorts in the earlier grades.

[22] Robert Coldiron, Pennsylvania State Department of Education, personal communication, January 1985.

[23] California State Assessment Program, unpublished tabulations.

Has the Upturn Ended?

The most recent National Assessment of reading found that the average reading proficiency of nine-year-olds was largely unchanged between 1979 and 1983, while the achievement of older students continued to rise.[24] That is, the birth cohort of 1974 showed no gain over the birth cohort of 1970.

Given the cohort pattern evident in the achievement upturn, this pattern--if it appears on other tests and is maintained--suggests that the upturn is, for the moment, over in the youngest age groups and that it will end fairly soon in the higher grades (as the birth cohorts that were nine years old between 1979 and 1983 pass through the grades). Tests administered to eighth graders would be expected to level off in the 1983 to 1987 period, while scores of seniors would level off between 1987 and 1991.

Whether this leveling off is a general phenomenon, however, is unclear. No other national data are available to test it, and state-level data are inconsistent. The proportion of New York third-grade students passing the state reference points in mathematics and reading, for example, has been stable since the 1970 and 1971 birth cohorts (see Figure B-5 in Appendix B). On the other hand. average scoıes in the elementary grades in Iowa have continued to rise, even in the most recent (1984-1985) year of data (see Figure III-2). In the next several years, National Assessments will take place in other subject areas, which will provide nationally representative data indicating whether this leveling off is a general occurrence.

DIFFERENCES IN TRENDS AMONG TESTS

Recent achievement trends have varied greatly from one test to another. For example, comparisons of recent trends on the SAT, the ACT, and standardized tests given to high school juniors and seniors as a whole show many discrepancies from one test to another (see Table III-4). This variation indicates that no single test, taken alone, is an adequate indicator of overall achievement trends. Indeed, in the absence of a clear understanding of the variations in the trends from one test to another, even a few tests taken together cannot always be assumed to be a sufficient indicator.

[24] National Assessment of Educational Progress, *The Reading Report Card: Progress Toward Excellence in Our Schools* (Princeton: NAEP/Educational Testing Service, 1985).

This variation apparently reflects differences both among the tests themselves and among the students taking them. The precise role of each is unclear, however, and some of the specific differences between tests are hard to explain. For example, although the ACT sample is in some important respects comparable to the SAT group and underwent some of the same compositional changes as affected the SAT, the trends on the two tests are markedly different.25/ Conversely, although the Iowa ITED is substantively similar to the ACT and was presumably free of many of the compositional changes that biased the SAT and ACT trends, it showed total declines roughly as large as those shown by the SAT (see Table III-2).

Subject Areas

The relative severity of the decline in different subject areas has been the focus of considerable discussion, in terms of both explanations of the trends and debates about appropriate responses. Debate has focused not only on specific subject areas, but also on two broad categories of subjects: those primarily taught "directly" in school, and those that are to a substantial degree taught "indirectly," both in school and elsewhere.26/ Some people would argue, for example, that certain mathematical skills--such as converting fractions to decimals or solving algebraic equations--are taught primarily in school through formal instruction and drill. In contrast, a larger proportion of vocabulary knowledge is presumably learned as an incidental result of daily experience at home and elsewhere. For this reason, a larger decline in the "indirectly" taught subjects might imply that the decline was attributable more to changes in student characteristics or to broad social changes than to changes in schooling, while larger declines in "directly" taught subjects would implicate schooling. 27/

25. Compositional changes affecting ACT means are discussed in Munday, *Declining Admissions Test Scores*.

26. Donald Rock and others, *Factors Associated with Decline of Test Scores*, p. 6.

27. While few people would argue with the idea that students learn a larger proportion of their vocabulary than of their mathematical skills outside of school, the observed relationships between achievement in different subject areas and home and school characteristics are not clear-cut. For example, an analysis of the relative size of home and school effects on achievement in several countries found that schooling effects were indeed larger in science than in reading among 10-year-olds but not among 14-year-olds (James S. Coleman, "Methods and Results in the IEA Studies of Effects of School on Learning," *Review of Educational Research*, vol. 45, Summer 1975, pp. 355-386, Tables 2 and 3.)

TABLE III-4. RECENT TRENDS ON STANDARDIZED TESTS AMONG HIGH SCHOOL SENIORS AND JUNIORS, WITH TRENDS OVER THE SAME PERIODS ON THE SAT a/

Test	Subject	Change (Standard Deviations)
	1970 to 1983	
National Assessment	Reading	.10
SAT	Verbal	-.26
ACT	English	.02
ITED-Iowa Grade 12	Vocabulary b/	-.08
	Reading c/	-.12
SAT	Mathematics	-.14
ACT	Mathematics	-.23
ITED-Iowa Grade 12	Mathematics b/	-.03
	1971 to 1979	
NLS to HSB	Vocabulary	-.22
	Reading	-.21
SAT	Verbal	-.26
NLS to HSB	Mathematics	-.14
SAT	Mathematics	-.15
	1970 to 1981	
Illinois Decade Study d/	English 1	-.38
	English 2	-.49
SAT	Verbal	-.25

(Continued)

TABLE III-4. (Continued)

Test	Subject	Change (Standard Deviations)
	1970 to 1981 (cont'd.)	
Illinois Decade Study	Mathematics 1	-.05
	Mathematics 2	-.26
SAT	Mathematics	-.14

SOURCES: CBO calculations based on National Assessment of Educational Progress, *The Reading Report Card;* Albert Beaton, NAEP/Educational Testing Service, personal communication, December 1985; Hunter M. Breland, *The SAT Score Decline;* Table 1, College Board, *National College-Bound Seniors, 1978 and 1985;* L. A. Munday, *Declining Admissions Test Scores;* and American College Testing Program, *National Trend Data for Students Who Take the ACT Assessment;* Iowa Testing Programs, "Mean ITED Scores by Grade and Subtest for the State of Iowa: 1962-Present;" Robert Forsyth, Iowa Testing Programs, personal communication, August, 1984; Donald A. Rock, Ruth B. Ekstrom, Margaret E. Goertz, Thomas L. Hilton, and Judith Pollack, *Factors Associated with Decline of Test Scores of High School Seniors, 1972 to 1980* (Washington, D.C.: Center for Statistics, U.S. Department of Education, 1985); *Student Achievement in Illinois, 1970 and 1981* (Springfield: Illinois State Board of Education, 1983).

a. The dates used in each set reflect the longest portion of the 1970-1983 period for which data are available. The NLS/HSB and Illinois Decade data are available only for the periods indicated. Comparisons extending past 1979 generally include a period of increasing scores.

b. These small changes in the ITED reflect substantial declines that were nearly offset by gains since 1978 and 1979.

c. This reflects the "Interpretation of Literary Materials" test. Reading is also tested on other tests in the ITED battery.

d. High school juniors only. SAT comparisons are therefore one year later.

Among the tests assessed here, no single subject area consistently showed the largest drop, and the decline was not consistently larger among either directly or indirectly taught subjects. In a majority of the tests, the drop was largest on language-related tests such as verbal reasoning, language usage, vocabulary, and reading. The exceptions were frequent enough, however, to suggest that this pattern is more reflection of the particular tests than an underlying characteristic of the achievement decline.[28] Indeed, a different assortment of tests--if more were available--might show a very different aggregate ranking of the decline in different subject areas.

Thus, for example, language-related tests showed the largest drops on the SAT, a nationally representative comparison of high school seniors in 1971 and 1979 (the NLS and HSB comparison), and in some of the Iowa data. Conversely, mathematics showed the steepest decline in other Iowa data and in the national normings of the California Achievement Test. Moreover, some of the language-related tests that showed particularly large declines (such as the vocabulary test in the NLS and HSB comparison) tap indirectly taught subjects, while others (such as the language test in the ITBS data and the ITED expression test) are clearly much more reliant on formal instruction. (For more detail on the relative size of the decline in different subject areas, see Appendix C.)

Underlying this seeming lack of consistency is the fact that achievement in any one subject can be defined--and measured--in many different ways, and the variations in measurement can be large enough to create very different trends. Thus, to speak of "the decline in mathematics achievement" is misleading. It is more accurate to speak of the decline in the mathematics skills measured by a specific test, and one should bear in mind that other tests might yield very different trends.

Trends in average mathematics achievement of Iowa students clearly illustrate the effect of test differences on the severity of the decline.[29]

[28] This discussion reflects only tests for which standard deviations are available, since the trends in different subject areas are made comparable by expressing them as fractions of a standard deviation. The National Assessment is therefore excluded, since standard deviations from previous assessments were not all retained by the NAEP staff. (Lawrence Rudner, Office of Educational Research and Improvement, U.S. Department of Education, personal communication, January 1985).

[29] Since most students in Iowa are tested with the ITBS through grade 8 and with the ITED in grades 9 through 12, differences between trends in Iowa on the grade 8 ITBS and the grade 9 ITED reflect little other than the differences in the tests themselves. The scores are based on almost the same group of students at nearly the same point in their school careers.

Over the entire period of the decline, the eighth-grade Iowa ITBS dropped substantially more in mathematics than in other subjects. In contrast, the ninth-grade Iowa ITED showed somewhat less decline in mathematics than in social studies or reading (the "interpretation of literary materials"). Over the whole period, the mathematics decline on the grade-eight ITBS was nearly half a standard deviation, or about .036 standard deviations per year. On the ITED, the total mathematics decline and the annual rate of the decline were both roughly half as large (see Figure III-5). The explanation of this difference might lie in the construction of the tests; the ITBS is roughly split between concept items (which are highly curriculum bound) and applications items, while the ITED places much greater emphasis on the latter. 30/

Level and Type of Skill

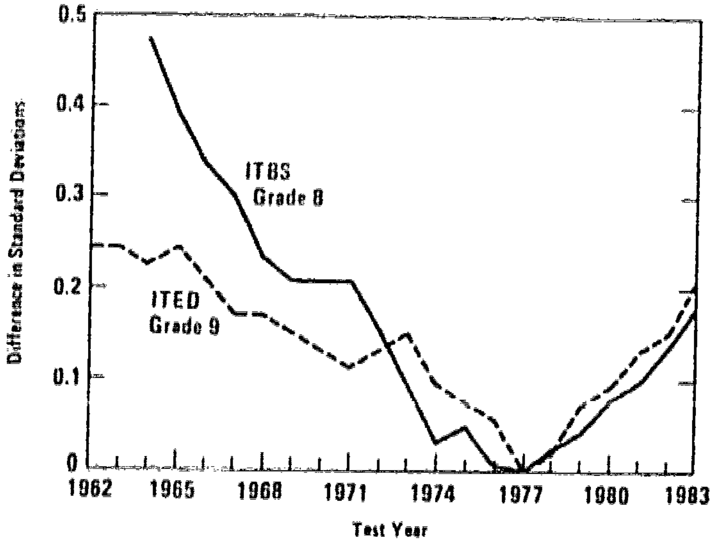
Evidence of the trends in different types and levels of skills is of two types: direct comparisons of different items within individual tests, and indirect inferences from comparisons of different tests--such as those in different subject areas or given at different grade levels. Direct comparisons of items can be carried out on any test, but little such analysis is currently available. Indirect inferences are therefore also noted in this section.

The Decline. That the overall drop in achievement entailed sizable declines in higher-level skills, such as inference and problem-solving, is beyond question. 31/ The extent to which declines occurred in more basic skills, such as simple arithmetic computation, is less clear. While some tests showed substantial declines in basic skills, other indices of basic skills showed little or no drop. In the aggregate, the evidence suggests that declines in the more basic skills might have been generally less severe than in higher-order skills, but not without exception.

[30] Robert Forsyth, Iowa Testing Programs, personal communication, February 1985.

[31] While the evidence leaves no doubt that substantial declines occurred in some higher-level skills, not all higher-level tests showed declines. The most notable exception is the Project TALENT 15-year retests, which showed increases in abstract reasoning and creativity in grades 9 through 11 between 1960 and 1975 (Table C-1 in Appendix C). This exception, however, might be artifactual. The starting point of the comparison --1960--antedated the predecline peak in achievement, thus confounding earlier growth in achievement with the decline (Cleary and McCandless, *Summary of Score Changes (in Other Tests))*. In addition, the 15-year retest suffers from two serious threats to validity and representativeness: a very small sample (only 17 schools), and meager assessment of changes in school characteristics that might bias the results.

Figure III-5.
Iowa Mathematics Achievement, Differences from Lowest Year



SOURCES: CBO calculations based on "Mean ITED Test Scores by Grade and Subtest for the State of Iowa" and "Iowa Basic Skills Testing Program, Achievement Trends in Iowa: 1955-1985" (Iowa Testing Programs, undated and unpublished tabulations); Robert Forsyth, Iowa Testing Programs, personal communication, April 1984; and A. N. Hieronymus, E. F. Lindquist, and H. D. Hoover, *Iowa Tests of Basic Skills: Manual For School Administrators* (Chicago: Riverside, 1982).

A greater decline in higher-order skills is apparent in the performance of 17-year-olds on the first two NAEP mathematics assessments (1972-3 and 1977-8), which span the last years of the decline. Performance on these tests was tabulated separately for four types of skills:

- o Knowledge: "recall of facts and definitions," including facts of the four basic arithmetic operations and measurement.

- o Skills: "the ability to use specific algorithms and manipulate mathematical symbols." This domain includes "computation with whole numbers, fractions, decimals, (and) percents...; taking measurements; converting measurement units; reading graphs and tables; and manipulating geometric figures and algebraic expressions." 32/

o Understanding: items "implying a higher level of cognitive process than simply recalling facts or using algorithms." Items in this domain required explanation or illustration of various skills and "transformation" of mathematical knowledge.

o Applications: items requiring the use of the preceding three types of skills, usually in problem-solving. 33/

Average performance in the simplest domain--mathematical knowledge--did not change at all during the five-year interval (see Table III-5). (An increase in performance on items involving metric measures offset a relatively small decline in the rest of this domain.) Both of the two highest levels--understanding and applications--showed declines. Moreover, the average performance in the applications domain was very low in both years. 34/ The "skills" area showed a comparably large decline, but within that area, the drop tended to be largest on the more complex items. 35/

The second international mathematics assessment by the International Association for the Evaluation of Educational Achievement (IEA) yielded results in grade eight that are comparable to the NAEP in this respect. Average achievement in grade eight fell over the 18 years between the first and second assessments, but the declines were greater "for more demanding comprehension and application items than they were for computation items." 36/ On the other hand, the same assessment found precisely the

32. At the simple pole, the "skills" domain incorporates items that would be considered "basic skills" by all observers--for example, simple arithmetic operations. At the other pole, it subsumes some fairly complex operations, such as solving a system of linear equations for *x* and *y* and solving quadratic equations.

33. National Assessment of Educational Progress, *Changes in Mathematical Achievement, 1973-78*, p. xi.

34. *Ibid.*, pp. 12-15.

35. *Ibid.*, pp. 4-9.

36. F. Joe Crosswhite, John A. Dossey, Jane O. Swafford, Curtis C. McKnight, Thomas J. Cooney, and Kennth J. Travers, *Second International Mathematics Study: Summary Report for the United States* (Champaign: Stipes Publishing Co., 1985), p. xi. Given the timing of the two assessments and the age of the students, the eighth-grade trends in the international assessment probably combine several years of increasing achievement with a longer, previous period of decline.

TABLE III-5. NAEP MATHEMATICS CHANGES 1972-1977, AGE 17, BY AREA (Average percent of items correctly answered)

Area	1972	1977	Change
Total	52	48	-4a/
Knowledge b/	63	63	0
Knowledge c/	63	62	-2a/ d/
Skills	55	50	-5a/
Understanding	62	58	-4a/
Applications	33	29	-4a/

SOURCE: NAEP, *Changes in Mathematical Achievement, 1973-78*, Tables 1-4, 6, and 7.

a. Statistically significant, *p* less than .05.

b. Including metric measures.

c. Excluding metric measures.

d. Components do not yield stated change because of rounding.

opposite pattern among 12th-grade students: an increase in achievement, much of which "was seen in the more demanding comprehension questions and, for the calculus students, at the even more demanding application level."37/ The 12th-grade results, however, were in large part a reflection of the performance of calculus students, who constitute a small and select segment of the senior class and whose performance may therefore say little about that of high school students in general.

Evidence of a greater decline in higher-order skills also appeared in the NAEP reading assessments. As noted earlier, 17-year-olds showed little total change in reading between the 1970-71 and 1979-80 assessments. The small (and statistically insignificant) change in total reading performance,

37. *Ibid.*, p. xi. Whether these gains reflect favorable trends during the period of the general achievement decline, an earlier or particularly sharp upturn, or both remains unclear. Because the final test was administered only a few years after the end of the general decline among seniors, however, it suggests that progress during the years of general decline played a role.

however, masks a somewhat larger (and statistically significant) decline in inferential comprehension (see Table III-6). As defined in the NAEP, inferential comprehension can be considered the highest-level skill tapped by the test. It entails comprehending ideas that are not explicitly stated by drawing inferences from material that is explicit. 38/ In contrast, literal-comprehension scores changed by only a trivial amount, and reference skills--also a more basic area--actually improved, albeit by a very small and statistically insignficant amount.

Deterioration of higher-level skills is also apparent from declines on tests that are designed specifically to tap them. 39/ The SAT is the most salient example. As noted earlier, it is designed (and is generally considered) to be a test that relies heavily on skills such as reasoning, problem-solving abilities, and verbal relationships (such as are assessed by analogies). The Illinois Decade Study (which used a test that was also developed by the Educational Testing Service) provides another example. While the Decade Study included many items that required that students know specific pieces of information (such as rules of English usage, social-studies facts, and mathematical terminology), it also relied heavily on inference. 40/ The declines on the test were relatively large (see Table III-4 and Appendix C).

The relationship between age and the size of the decline--discussed in Chapter IV--might also be indirect evidence of a lesser deterioration of more basic skills. As noted earlier, declines in the first three grades tended to be slight and short-lived and might best be seen as brief interruptions of an otherwise steady upward trend in those grades. Since the curriculum in

38. NAEP, *Three National Assessments of Reading*, pp. 4, 25.

39. The tests noted here are all multiple-choice format. As noted in Chapter 2, some people have argued that multiple-choice tests are demonstrably limited in their ability to tap many higher-order skills (for example, Norman Frederiksen, "The Real Test Bias: Influences of Testing on Teaching and Learning," *American Psychologist*, vol. 39 (March 1984), pp. 193-202). Even if the tests noted here leave many relevant higher-order skills unassessed or inadequately measured, however, few people would argue with the notion that they do rely substantially on some higher-order abilities and that those abilities play a greater role in determining scores in these tests than in some others (such as the NAEP literal comprehension reading subtest or the NAEP mathematics test as a whole).

40. Illinois State Board of Education, *Student Achievement in Illinois, 1970 and 1981*, Appendix A.

TABLE III-6. NAEP READING CHANGES 1970-1979, AGE 17, BY AREA (Average percent of items correctly answered)

Area	1970	1979	Change
Total	68.9	68.2	-0.7
Literal Comprehension	72.2	72.0	-0.2
Inferential Comprehension	64.2	62.1	-2.1[a]
Reference Skills	69.4	70.2	0.8

SOURCE: NAEP, *Three National Assessments of Reading: Changes in Performance, 1970-80*, Table 6.

a. Statistically significant, *p* less than .05.

the early grades includes a large amount of basic skills--decoding and literal comprehension in reading, memorization of basic arithmetic facts, learning of the simplest arithmetic algorithms, spelling, and so on--the almost uninterrupted progress in those grades might reflect relatively favorable trends in the mastery of those particular basic skills.

The Subsequent Upturn. The characteristics of the subsequent upturn are as yet less clear, in part because scores on tests administered in high schools began improving only recently. That the upturn is occurring in most tests and at all grade levels--including the SAT in the last few years--suggests that improvements are probably occurring at many skill levels, but there is as yet little direct indication of the relative size of the upturn in different types and levels of skills. Moreover, the pattern may be complex; for example, the upturn may have different components in different grade levels or among different groups of students.

Disquieting but incomplete suggestions of relatively smaller increases in higher-level skills are found in the most recent (1981-82) NAEP mathematics assessment. Because the NAEP tests a nationally representative sample of students and because it permits comparisons of changes in various skill areas, it is a particularly important indicator of the mix of skills comprised by recent trends. The NAEP found a sizable increase in the performance of 13-year-olds between 1977-1978 and 1981-1982 (but no appreciable change in the performance of 17-year-olds and only slight and statistically insignificant gains among 9-year-olds). 41/ The nature of the improvement among 13-year-olds, however, was disturbing:

> ...They improved most on the knowledge, skills, and understanding exercises, and least on the applications exercises. Further study shows that their improvements in understanding came on exercises judged relatively easy by a panel of mathematics [illegible]rs; performance levels on exercises calling for [illegible]r understanding showed little or no improvement. 42/

On the other hand, recent gains among the highest-achieving students on difficult tests--discussed in the following chapter--suggest improvement in their higher-order skills. It is possible that some groups of the highest-achieving students are gaining substantially in higher-order skills, while many other students are showing less progress in this regard, but available data remain too limited to answer this question.

41. The lack of change among 17-year-olds, but not the absence of substantial improvement among 9-year-olds, is predictable on the basis of the cohort model discussed earlier.

42. NAEP, *Third National Mathematics Assessment*, p. xv.

CHAPTER IV

GROUP DIFFERENCES IN ACHIEVEMENT TRENDS

While the achievement decline and the subsequent upturn occurred among most groups of students identifiable in the existing data, both trends varied among different groups. Similarly, achievement trends have varied among different types of communities and schools.

The most important differences in trends are:

- Greater declines on tests administered to older students;
- Relative gains by black and Hispanic students, compared with nonminority students; and
- Relative gains in high-minority schools and schools in disadvantaged urban communities compared with the nation as a whole.

In addition, there is some indication that students in the bottom fourth of the achievement distribution gained ground relative to those in the top fourth during part of the 1970s. The evidence on this point is inconsistent, however, and it is not clear that this narrowing of the gap occurred on a variety of tests or spanned more than a short period of time. Female students also showed slightly sharper declines on language-related tests (such as reading and vocabulary), but not on tests in other subject areas. Private school students showed declines comparable to those among public school students in reading and vocabulary, although evidence from a single test suggests that the decline in mathematics achievement was considerably smaller among private-school students.

DIFFERENCES IN TRENDS AMONG TYPES OF STUDENTS

Variation in achievement trends were associated with age, sex, achievement subgroup (that is, low versus high achievers), and race and ethnicity.

Age

Both the decline in achievement and the subsequent upturn varied markedly with the age of the students tested, but the effects of age appear to have been different during the two periods.

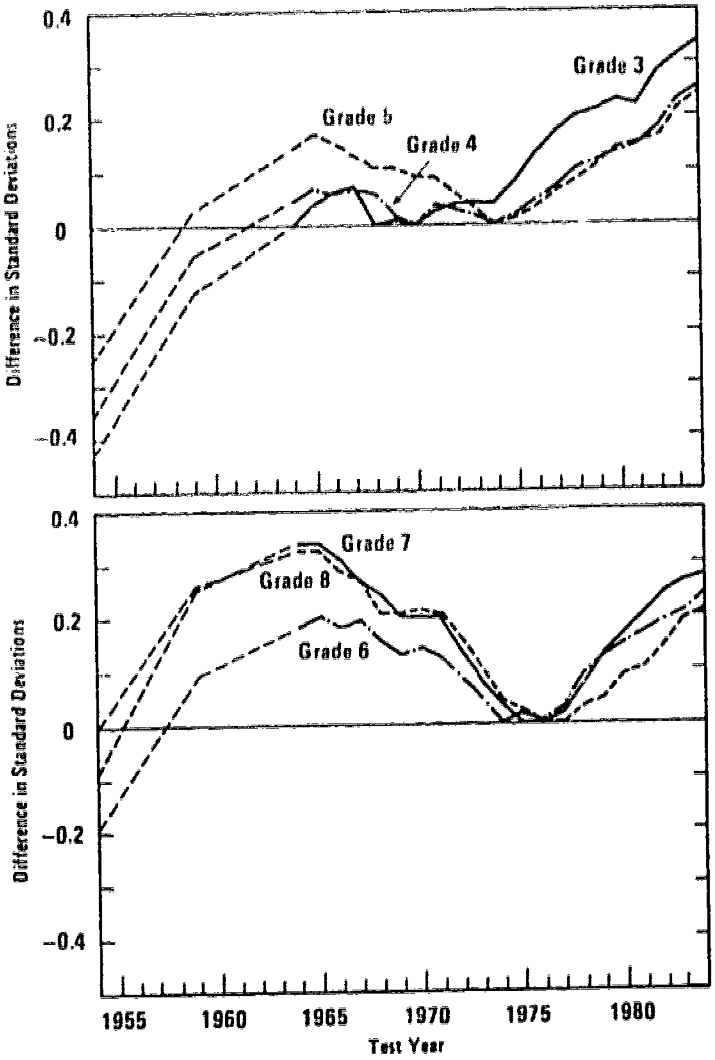
The Decline. The total size of the decline was strongly related to age. In general, tests administered to older students showed markedly larger total declines than did tests administered in the early grades. 1/

The Iowa state data provide the best assessment of this question and show a striking link between age and the size of the achievement decline (see Figures IV-1 and IV-2). 2/ At one extreme, the decline in third-grade scores was small and short-lived; it can be characterized as a slight dip accompanying an eight-year hiatus in an otherwise unbroken, 30-year increase in achievement. In standardized form, the total decline was only about 0.07 standard deviation (depending on subject), and average scores are now over a third of a standard deviation above the low point of the decline--and more than three-fourths of a standard deviation above their level of

1. Although this conclusion is widely accepted, it is important to note that it is actually based on fairly limited data. To offer a good test of the relationship between age and the size of the decline, a data series should meet a set of criteria that few do. The data series should include comparable tests administered to a range of ages, since a comparison of different tests can confound differences between the tests themselves with the effects of age. Scores should be presented in some form--such as standard deviations or percentiles--that permits comparisons among grades. The data should also extend back to the onset of the decline. Data that extend over a relatively short period of time might tap a relatively steep portion of the decline in one grade and a relatively gradual portion in another, thus biasing the comparison among age groups. In addition, random year-to-year fluctuations in scores--reflecting either sampling fluctuations or uncontrolled differences in tests--are more likely to bias conclusions based on a relatively few years. Finally, the data should be annual, to confirm that they subsume the entire decline and none of the upturn. Data that are collected intermittently--such as the NAEP and norming data from commercial elementary and secondary tests--can mix in varying periods of increasing scores for different age groups. Intermittent data also might capture a relatively steep portion of the decline in one grade but a comparatively gradual portion in another.

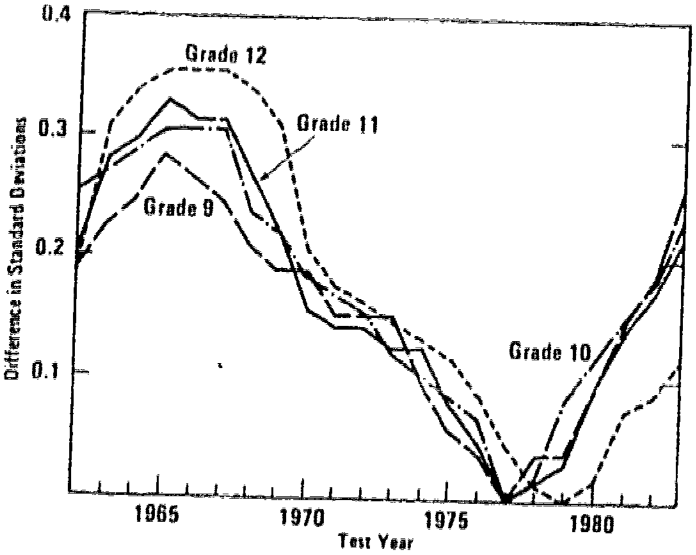
2. The best assessment of the effect of age is obtained within each test series--that is, comparing ITBS scores in grades 3 through 8 with each other, and similarly comparing ITED scores in grades 9 through 12. Even in this case, comparisons across the two tests --for example, comparing grade 8 ITBS scores with the grade 9 ITED--confounds differences between the two tests with the effects of age. (See the discussion in Chapter III of differences in trends among subject areas for a concrete example of differences of this sort between the ITBS and ITED.)

Figure IV-1.
Iowa Composite, ITBS, Grades 3-8, Differences from Post-1964 Low Point



SOURCES: CBO calculations based on "Iowa Basic Skills Testing Program, Achievement Trends in Iowa: 1955-1985" (Iowa Testing Programs, unpublished and undated material); and A. N. Hieronymus, E. F. Lindquist, and H. D. Hoover, *Iowa Tests of Basic Skills: Manual For School Administrators* (Chicago: Riverside, 1982).

Figure IV-2.
Iowa Composite, ITED, Grades 9-12, Differences from Lowest Year



SOURCES: 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); Robert Forsyth, Iowa Testing Programs, personal communication, August 1984.

three decades ago. Thus, the median third-grader in Iowa today scores better than roughly 68 percent of his or her counterparts of three decades past. Similarly, no sizable decline occurred in grade three in statewide assessments in New York and California. 3/

The decline in eighth-grade Iowa scores, in contrast, was large enough to depress composite achievement scores to their level of three decades ago and long enough that recovery has as yet been incomplete. When put in standard form, these differences appear even more striking. Eighth-grade Iowa scores declined about a third of a standard deviation and have since

3. New York State Education Department, unpublished tabulations; Frank Armbruster, Paul J. Bracken, and Joyce Lind, *The U. S. Primary and Secondary Educational Process* (Croton-on-Hudson, New York: The Hudson Institute, 1975), Appendix A; Dale Carlson, California State Department of Education, personal communication, March 1984.

recovered only about two-thirds of what they lost. (Nonetheless, eighth-grade scores are still about 0.2 standard deviation higher than they were 30 years ago, placing the median student this year at the 58th percentile relative to achievement levels in 1954.)

The National Assessment of Educational Progress (NAEP) also shows only relatively few and small declines among nine-year-olds, relative to the declines in the older groups. This pattern might in part reflect the timing of the NAEP assessments, however, rather than--or in addition to--truly lesser declines in the youngest age group. 4/

Periodic national norming data from commercial standardized elementary and secondary tests also suggest both a lack of decline in the youngest age groups and progressively larger declines in the remainder of the school-age population. For example, the national ITBS norming data indicate that in reading, the median third-grader's level of achievement increased by 4.3 months from 1955 and 1963, only 0.5 months from 1963 to 1970, and 3.7 months from 1970 to 1977. This change is consistent with the pattern in the annual Iowa data--that is, a pause in achievement growth in the late 1960s and early 1970s. In contrast, among sixth graders, a 2.2-month gain from 1955 to 1963 was followed by declines of 2.6 and 3.0 months in the following seven-year periods. Among eighth graders, the drop was even more substantial after 1970. 5/ The SRA achievement series showed composite gains in all but one grade between 1962 and 1971. Between 1970 and 1977, however, the trends varied greatly with grade level. In reading, for example, the latter period included large gains (two-thirds of a standard deviation or more) in grades one and two; more moderate gains in grades three and four; small declines in grades five through eight; and larger drops in the higher grades. 6/

4. Given the cohort pattern shown by the end of the decline, the various NAEP assessment cycles probably began near or even at the end of the decline among nine-year-olds, and thus the data most likely combine a few years of the decline with a longer period of the subsequent increase. Since the NAEP assessments are conducted only at intervals of four or five years, however, the precise end of the decline in that test series cannot be firmly established, and the extent of this confounding therefore cannot be determined.

5.. A. N. Hieronymus, E. F. Linquist, and H. D. Hoover, *Iowa Test of Basic Skills: Manual for School Administrators* (Chicago: Riverside Publishing Company, 1982).

6. Science Research Associates, *SRA Achievement Series, Technical Report #3* (Chicago: SRA, 1981), Table 2; and Science Research Associates, unpublished tabulations. The trends between the 1970 and 1977 school years reported here reflect normings conducted in the springs of 1971 and 1978 and are labeled in terms of those calendar years in the published data.

Although the achievement decline persisted longer in the higher grades, the larger total drop in scores in those grades reflects more than the longer duration. In addition, the decline appears to have been steeper--that is, more rapid--in the higher grades. This rapidity is shown most clearly by the Iowa data (both the ITBS and the ITED; see Figures IV-1 and IV-2). In all but two cases, the decline in any grade was steeper than that in all lower grades. This difference in the rapidity of the decline, however, appears to have been confined primarily to the earlier years of the decline.

The Upturn. As noted in Chapter III, scores on tests administered to younger children have risen substantially more in recent years, compared with the decline in those grades, than have scores on tests administered in the higher grades. This pattern can be seen clearly in the Iowa state trend data (both the ITBS and ITED; see Figures III-2 and III-3):

- Grades 3, 4, 5, and 6 are now at their highest point in the three decades of available data.

- Achievement in grades 7, 8, 9, and 10 has rebounded strongly but is not yet at its earlier high (although grade 9 is nearly at that level).

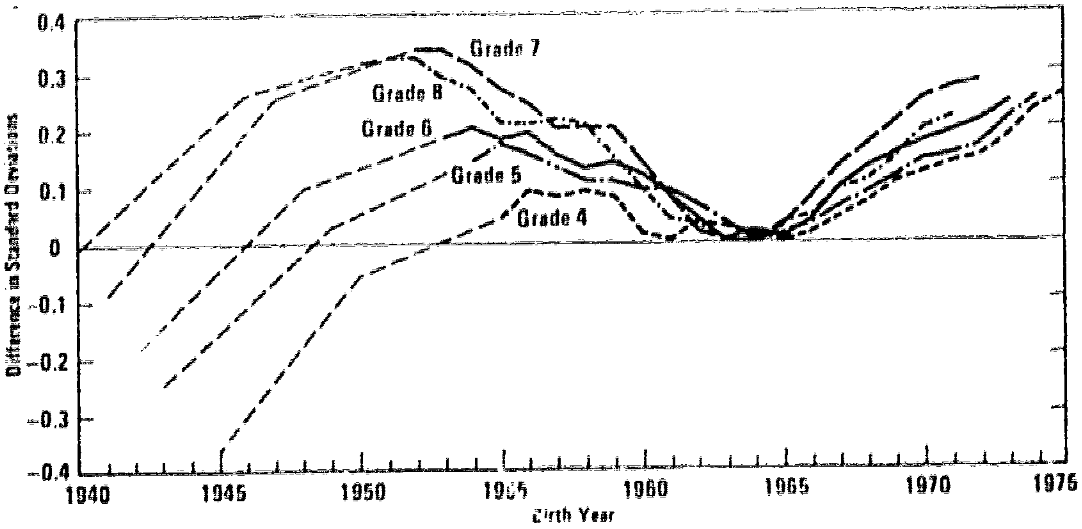
- Grade 12 achievement has begun rising but remains near its low point.

The well-known SAT 'rend parallels the twelfth grade Iowa trend in this regard: achievement has been climbing for several years but remains only modestly above its low point (see Figure III-4). Similar patterns--although often less clear-cut--appear in a number of otner data bases as well, such as the Virginia State assessment data and the NAEP reading assessment. (Some achievement test series, however, are inconsistent with this pattern. For example, in the NAEP mathematics assessment, the recent increase in performance was markedly greater among 13-year-olds than among 9-year-olds.) 7/

The greater total rise in scores to date in the younger grades appears largely to reflect a longer period of rising scores in those grades rather than a greater rate of improvement than in the higher grades. The upturn in scores followed quickly after the end of the decline and shows the same

7. National Assessment of Educational Progress, *The Third National Mathematics Assessment: Results, Trends, and Issues* (Denver: NAEP/Education Commission of the States, 1983), Table 5.1.

Figure IV-3.
ITBS Composite, by Birth Year



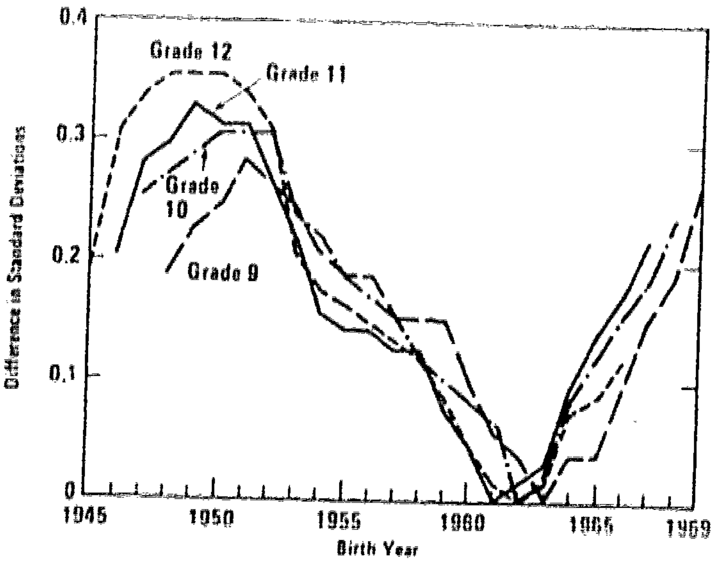
SOURCES: CBO calculations based on "Iowa Basic Skills Testing Program, Achievement Trends in Iowa, 1955-1985" (Iowa Testing Programs, unpublished and undated material); and A. N. Hieronymus, E. F. Lindquist, and H. D. Hoover, *Iowa Tests of Basic Skills: Manual For School Administrators* (Chicago: Riverside, 1982).

cohort pattern (see Appendix B). Among children born after 1963 or so--that is, beginning with the cohorts that entered school in the late 1960s--each cohort has tended to outscore those preceding it. The smaller gains in the higher grades thus appears to reflect, at least in substantial part, the smaller number of higher-scoring cohorts that have reached senior high school.

This trend can be seen in the Iowa data, which suggest--if trends in Iowa are indicative of national trends in this regard--that gains have been comparably fast, or even more rapid, in the higher grades than in the lower ones.[8] On the ITBS, each birth cohort since the onset of the score increase has tended to produce slightly larger increases in grades six through eight than in grades four and five (see Figure IV-3; vertically adjacent lines that

[8] This conclusion reflects changes expressed in standard deviations and only comparisons within a single test. Trends on the ITED are not compared with those on the ITBS.

Figure IV-4.
ITED Composite, by Birth Year



SOURCES: CBO calculations based on "Mean ITED Test Scores by Grade and Subtest for the State of Iowa: 1962 to Present" (Iowa Testing Program, undated and unpublished tabulations); Robert Forsyth, personal communication, August 1984.

are parallel indicate comparable gains by the same cohort in different grades). In this respect, the upturn in the ITBS has been largely symmetrical with the last years of its downturn. On the ITED, the gains produced by any given cohort have remained roughly comparable as that group moved from grade 9 through grade 12 (see Figure IV-4). For several cohorts after the upturn began, these gains were also basically symmetrical with the corresponding last years of the decline, but the most recent cohorts to reach the high-school years--those born in 1966 through 1969--have produced gains that are larger than the corresponding decline produced by the birth cohorts of the mid-1950s.

Sex

While the achievement decline was sizable among students of both sexes, it was somewhat more severe among female students in the case of language-related tests (such as vocabulary, reading, and the SAT-Verbal). On the other hand, once the effects of changes in the composition of the test-taking group are taken into account, the declines among males appear to have been comparable or even slightly larger than those among females in mathematics and science. 9/

The average SAT sc res of women dropped substantially more than those of men. This difference by sex was large on the verbal scale--after 1967, women dropped 50 points, compared with the 36-point drop in the average score of males--but far smaller on the mathematical scale. 10/ The average score of female ACT candidates also dropped more than that of males, and the difference was greater on the English test than in mathematics. 11/

In both cases, however, the apparently greater decline among women might simply be a reflection--at least in part--of the changing mix of male and female students taking the tests. Women have constituted a growing share of all students taking both the SAT and the ACT. Women constituted 42.7 percent of SAT candidates in 1960, 47.5 percent in 1970, and 51.8 percent in 1983. 12/ Similarly, women constituted 45 percent of ACT candidates in 1964 and 54 percent both in 1975 (the year that ACT scores

[9] On the ACT, the greater decline among women was most pronounced in social studies. In the NAEP, however, the only comparison in social studies that showed relatively greater trends in one gender than t e other--citizenship questions at age 13--showed females gaining relative to maler. Comparable tabulations from other tests are unavailable. The sharp decline of women on the ACT social studies test therefore might be just a reflection of the compositional changes discussed below. L. A. Munday, *Declining Admission Test Scores*, Research Report #71 (Iowa City: American College Testing Program, February 1976); National Assessment of Educational Progress, *Changes in Political Knowledge and Attitudes*, 1969-76 (Denver: NAEP/Education Commission of the States, March 1978.)

[10] College Entrance Examination Board, *College-Bound Seniors, 1984* (New York: The College Board, 1984).

[11] These patterns reflect changes in ACT scores from 1965 to 1975, the latter being the year in which composite ACT scores reached their lowest point. The data from 1965 to 1969 are slightly inconsistent with the later data because the former include residual on-campus testing. The former are taken from Munday, *Declining Admission Test Scores*; the latter are from unpublished ACT tabulations.

[12] Advisory Panel on the Scholastic Aptitude Test Score Decline, *On Further Examination* (New York: College Entrance Examination Board, 1977), p. 16; and College Board, *College-Bound Seniors, 1984*.

reached their low point) and in 1983. 13/ This growing share suggests that the pool of women taking the tests might have become relatively less select--a change that would lead to greater score declines among women than among men.

Trends in scores on other tests, however, suggest that part of the greater decline among women is independent of these compositional changes, reflecting some other, as yet unidentified, factors. Data from a few nationally representative tests--which are largely free of these compositional changes--also show greater declines among female students on language-related tests. 14/ On the other hand, in mathematics and science the decline in the scores of male students was typically as large or even larger. For example, a comparison of the high-school classes of 1972 and 1980 found that women showed a greater decline in vocabulary and a slightly larger drop in reading, while men showed a larger decline in mathematics. 15/ Seventeen-year-olds showed a similar pattern in the NAEP over a five- to nine-year span in the 1970s. Women showed a greater decline on both the literal comprehension and inferential comprehension components of the reading assessments, while men evidenced slightly greater declines in mathematics and science. 16/ Although these differences by sex were very small, they might have been larger if the comparisons had spanned the entire period of the achievement decline rather than only a portion of it.

13. Munday, *Declining Admissions Test Scores*; American College Testing Program, unpublished tabulations.

14. Although nationally representative data are most often largely free of this particular type of compositional change, they are not always entirely devoid of it. For example, data based on high school samples could show a change of this sort if trends in dropout rates differed markedly by sex.

15. Donald A. Rock, Ruth B. Ekstrom, Margaret E. Goertz, Thomas L. Hilton, and Judith Pollack, *Factors Associated with Decline of Test Scores of High School Seniors, 1972 to 1980* (Washington, D.C.: Center for Statistics, U.S. Department of Education, 1985).

16. National Assessment of Educational Progress, *Three National Assessments of Reading* (Denver: NAEP/ Education Commission of the States, 1981), Tables A-9, A-10, and A-11; *Mathematical Technical Report: Summary Volume* (Denver: NAEP/ Education Commission of the States, 1980), Table 4; *Three National Assessments of Science: Changes in Achievement, 1969-77* (Denver: NAEP/ Education Commission of the States, 1978), Table A-4. In the case of science, the scores of women increased, while those of men dropped.

Achievement Subgroups

A current and widely held view is that the decline in achievement was more severe among relatively high-achieving students than among those at the lower end of the achievement distribution. This belief has led some observers to credit the educational system with improving its services to low-achieving students, or, alternatively, to fault it for allowing its services for more able students to deteriorate. 17/

It is not clear, however, that trends have been consistently more favorable among lower-achieving than among higher-achieving students over the entire period of the achievement decline and subsequent upturn. When a wide range of tests is considered, a more complex--and sometimes inconsistent--pattern emerges. Moreover, there are major gaps in the available data--such as the sparseness of relevant comparisons during the first half of the achievement decline, and a very limited picture of the relative performance of achievement subgroups during the recent years of increasing achievement. In addition, both apparent changes in the gap between achievement subgroups and inconsistencies in the data about these groups must be taken cautiously because both consistencies and variations in the data can be artifacts of technical aspects of the tests.

As discussed in Chapter II, a number of technical aspects of tests influence conclusions about relative trends in high- and low-achieving groups. Differences in the scaling of test scores can markedly affect such judgments. In addition, a single test is unlikely to be a comparably comprehensive measure of mastery at two very different levels of achievement and therefore may understate the relative change of students at one level. The tabulation and reporting of results further complicates comparisons, since information on the additional items correctly or incorrectly answered is rarely reported, particularly for achievement subgroups. This lack of information makes it hard to judge whether changes in the average scores of achievement subgroups are substantively comparable, even when they seem similar numerically. Nonetheless, the broad range of tests suggests the following generalizations. (See Appendix D for additional details.)

17. See, for example, statement by Archie E. Lapointe, Executive Director, National Assessment of Educational Progress, before the Subcommittee on Elementary, Secondary, and Vocational Education, Committee on Education and Labor, January 31, 1984; and William W. Turnbull, *Changes in SAT Scores: What Do They Teach Us?* (report to the College Board-ETS Joint Staff Research and Development Committee, forthcoming).

It is clear that the achievement decline and the subsequent upturn appeared among both low- and high-achieving students. Whether the decline began at the same time in different achievement subgroups, however, and whether the drop was comparable among those subgroups during the early years of the decline (the late 1960s and the first years of the 1970s) remain unknown. Tabulations comparing achievement subgroups during those years are largely restricted to unrepresentative groups of students--for example, comparisons of students taking the SAT, classified in terms of their rank on that test.

During the mid- and late 1970s--that is, during the end of the achievement decline and the beginning of the subsequent upturn--students in the top achievement quartile (the top fourth of all students, when ranked by achievement) lost ground relative to those in the bottom quartile in reading, mathematics, and science in the National Assessment of Educational Progress. That pattern appeared in all three age groups tested (ages 9, 13, and 17), although it took different forms at different ages--probably as a result of the cohort pattern shown by the end of the decline. At age nine, gains predominated over losses, but the lowest quartile showed larger gains than did the highest. At age 17, declines predominated, with the larger losses generally appearing in the highest quartile. At age 13, gains and losses were more evenly mixed, but the lowest quartile still showed relative gains.

While the narrowing of the gap between the top and bottom achievement quartiles on the NAEP is clear-cut, other data cast doubt on the extent to which this was a general trend over the past two decades. Similar trends appear in some data (such as the Illinois Decade Study and some tabulations of the SAT), but not on others (such as other tabulations of the SAT).[18] Moreover, under most circumstances, a narrowing of the gap between the top and bottom quartiles would cause the standard deviation of test scores--that is, their variability--to decrease. That has not been the general pattern, however, in the few data sources for which historical records of standard deviations are available. Since the early 1970s, the standard deviations (SDs) of the SAT and ACT have been stable or increasing slightly. The SD of the ITBS has been increasing, while that of the SRA achievement series has shown mixed trends (generally inconsistent with

[18] The Illinois Decade Study is a comparison of the performance of Illinois high school juniors on a fairly high-level battery of achievement tests in the 1970 and 1981 school years. See Appendix D.

the NAEP pattern in the earlier grades, but consistent in the higher grades). 19/

Several explanations of this inconsistency are plausible. Some of the variation among tests could simply be an artifact of scaling differences. For example, the Illinois Decade Study is consistent with the NAEP in its published form, which presents simple differences in scores, but is inconsistent when presented in terms of proportional changes in scores. Differences in the way students are classified as high- and low-achieving could also account for much of the variability. For example, classifying students in terms of their self-reported class rank yields patterns on the SAT since 1975 that are consistent with the NAEP (even though the standard deviation of the SAT was increasing at that time), while classifying students in terms of their rank on the SAT itself yielded trends that are inconsistent with the NAEP. On the other hand, some of the inconsistency might reflect true variation among tests; perhaps the lowest quartile gained relative to the highest only on certain types of tests.

Test scores of students taking college admissions tests--currently, about half of all high school graduates--declined more than those of high school seniors in general. But this difference primarily reflects the changing composition of the group taking those tests rather than a greater decline in achievement among high-achieving students. The proportion of students taking the SAT, for example, grew substantially during the 1960s and early 1970s, and this growth was accompanied by an increase in the share of SAT candidates from historically lower-achieving groups, such as certain ethnic groups and families of lower socio-economic status.20/ Since the early 1970s, however, such changes in the composition of the test taking group have been relatively minor. 21/

The highest-achieving students--those scoring highest on tests, taking the most advanced courses, and so on--evidenced both the decline and the subsequent upturn in achievement. These students did not show a consistently greater decline than the average student. Indeed, by some measures,

[19] The College Board, *College-Bound Seniors*, various years; American College Testing Program, unpublished tabulations; H.D. Hoover, personal communication, March 1984; and Science Research Associates, *SRA Achievement Series, Technical Report #3*, Table 2.

[20] Advisory Panel on the Scholastic Aptitude Test Score Decline, *On Further Examination*.

[21] Because compositional changes exacerbated the decline in the SAT but not the subsequent upturn, comparing the SAT upturn to the previous decline is misleading. The relative size of the upturn is understated unless adjustments are made to compensate for the compositional changes.

they appear to have gained recently relative to the average, particularly in the area of mathematics. For example, the proportion of SAT candidates scoring over 700 on the mathematics test has risen sharply in the last few years (from 2.7 percent in 1980 to 3.6 percent in 1984) and is now quite close to the level of 1966--the highest level in any year for which tabulations are available. Similarly, American seniors taking calculus and pre-calculus--together about 10 percent to 12 percent of all seniors--showed gains between 1964 and 1981 in international assessments of mathematics achievement. The sketchiness and inconsistency of data on the highest-achieving students, however, cloud these conclusions.

RACE AND ETHNICITY

Recent years have seen a shrinking of the long-standing difference between the scores of black and nonminority students on a variety of achievement tests. The evidence pertaining to other ethnic groups is more limited, but there are suggestions of relative gains by Hispanic students as well.[22] While the change has been small relative to the remaining gap between the minority and nonminority students, it has been consistent from year to

[22] The term "ethnicity" as used in the following discussion encompasses some distinctions --such as that between blacks and whites--that are often popularly termed racial. This convention is followed in part for simplicity, but also because some of the most common current categories have at best ambiguous racial bases. For example, many South Asians are often classified as nonwhite (as in some Census tabulations), even though most South Asians are in fact racially Caucasian. Similarly, people of mixed black/white origin are frequently classified as black without regard to whether the greater proportion of their ancestry is in fact white or black. Hispanics are almost all classified as whites in Census tabulations, even though many of them are racially mixed. (In particular, many are partially or primarily native American in origin, and native Americans are racially classified as "Mongoloid"--that is, Asian--people.)

The ethnic categories used in this paper necessarily reflect the disparate conventions used in the data sources cited and therefore vary among tests. In general, the term "nonminority" excludes, to the extent possible, all minority groups identified in each data source and usually corresponds to the category labeled "white" in the cited sources. The data sources vary considerably, however, in terms of how many--and which--groups are specifically identified. Moreover, some individuals--such as black Hispanics--can be classified in more than one way, and there is typically little information available about how those ambiguous cases are handled.

The more important known variations in the classifications used in the various sources are noted in Appendix E.

year and could prove substantial over the long run. These patterns are summarized below and are discussed in more detail in Appendix E.

Trend data on the scores of different ethnic groups are very limited, however, and generally extend back only a relatively short length of time. In addition, since many ethnic-group differences in achievement are large, the ambiguity inherent in measuring changes in the gaps between achievement subgroups described above applies to these comparisons as well. In this case, however, the pattern of the trends leaves no doubt that the closing of the gap is at least in part real and not an artifact of the tests.[23] Finally, classification of students' ethnicity is likely to be prone to error, both because of the u:.reliability of students' self-reports and because of the ambiguity--and lack of consistency over time--of ethnic classifications. While this is unlikely to be a serious source of bias in interpreting trends among black students, it is cause for caution in considering data about Hispanics.[24]

Black Students. In general, it appears that the average scores of black students:

o Declined less than those of nonminority students during the later years of the general decline;

o Stopped declining, or began increasing again, earlier; and

o Rose at a faster rate after the general upturn in achievement began.

[23] This narrowing of the gap is substantiated by several factors. First, the pattern is consistent among a variety of very different tests. Second, during certain periods, the convergence reflected gains among blacks concurrent with declines among nonminority students. Unlike differences in relative gains (or declines) between groups, a pattern of gains in one group and declines in the other is unlikely to be an artifact of the scaling method used and will generally persist even if the data are rescaled. Third, biases caused by ceiling effects have been largely ruled out. In the case of tests scored as the percent of questions answered correctly, the scores of the higher-achieving group can be held down by a ceiling effect, creating an illusion that lower-achieving groups are gaining in comparison. To lessen the likelihood of such a distortion, data of that sort were transformed (by a logit transformation) to eliminate ceilings, and the narrowing of the gap remained.

[24] See, for example, "Problems in Defining Ethnic Origin," Appendix A in Congressional Research Service, *Hispanic Children in Poverty* (Washington, D.C.: CRS, September 13, 1985).

The relative gains of black students appear on a variety of tests administered to students of different ages in different localities. They appear at ages 9, 13, and 17 in the National Assessment of Educational Progress (Figure IV-5); in the SAT; in a nationally representative comparison of high school seniors in 1971 and 1979; in grades 3, 6, and 9 in the North Carolina state assessment program; among ninth graders in the Texas state assessment program; and in test data from some local education agencies, such as Cleveland, Houston, and Montgomery County (Maryland). 25/

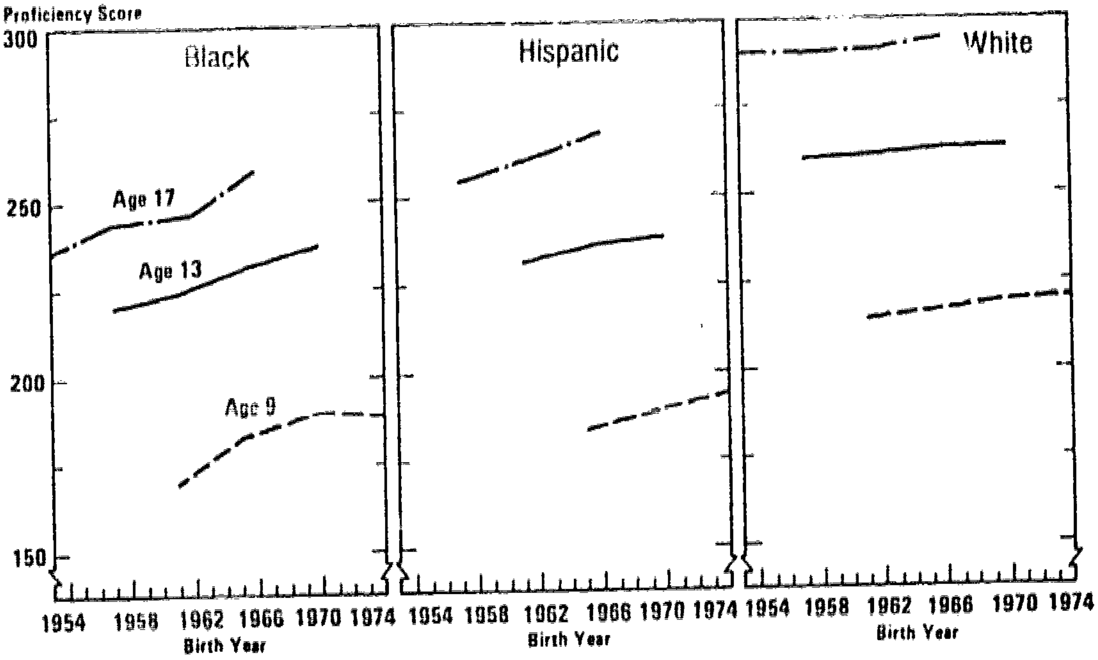
The SAT data suggest that part of the convergence of black and nonminority scores resulted from the decline ending earlier among black than among nonminority students. The convergence of scores continued during the period of the general upturn, however, as black students gained more rapidly than did nonminority students.

Although this shrinking of the gap has been small relative to the average differences between black and nonminority students, the rate of change has been appreciable. For example, over the past nine years, the gap between black and nonminority students on the SAT has shrunk at an annual rate roughly comparable to the average rate of the total SAT decline--a change that few people would label insignificant. On the National Assessment, the average black student's mathematics score was a third below the nonminority average in 1972 but a fourth below that in 1981.

25. National Assessment of Educational Progress (NAEP), *Three National Assessments of Science;* NAEP, *Three National Assessments of Reading;* NAEP, *The Reading Report Card;* NAEP, *The Third National Mathematics Assessment;* and NAEP, *Mathematical Technical Report: Summary Volume; College Board Data Show Class of '85 Doing Better on SAT, Other Measures of Educational Achievement* (New York: The College Board, September 23, 1985); Rock and others, *Factors Associated with Decline of Test Scores,* Tables D-1, D-2, and D-3; Nancy W. Burton and Lyle V. Jones, "Recent Trends in Achievement Levels of Black and White Youth," *Educational Researcher*, vol. 11 (April 1982), pp. 10-14, 17; Montgomery County (Maryland) Public School District, "MCPS Test Results by Racial/Ethnic Groups, 1977-1982," unpublished paper; Marian Kilbane-Flash, personal communication, March 1984; and Houston Independent School District, unpublished tabulations.

On the other hand, scores on the ACT are not entirely consistent with this pattern. The gap between black and other students on the ACT composite has narrowed since 1970, but only slightly, and the trend has been highly erratic from year to year. In addition, the trend varies among subjects; the gap narrowed in social studies, for example, but grew slightly in mathematics. This partial inconsistency with the patterns evident in other tests is discussed further in Appendix E.

Figure IV-5.
Trends in Average Reading Proficiency for White, Black, and Hispanic Students, by Birth Year



SOURCE: National Assessment of Educational Progress, *The Reading Report Card* (Princeton: NAEP/Educational Testing Service, 1985), Data Appendix.

It is likely, but not certain, that this narrowing of the gap will continue to appear in some test data for several years. The NAEP data show the most rapid convergence accompanying the birth cohorts of the mid-1960s as they pass through school--appearing at age 9 in the early 1970s and at age 17 in the early 1980s. Some narrowing, however, appeared at least as late as the birth cohorts of the late 1960s and perhaps as late as those of the early 1970s. 26/ This pattern would suggest further convergence between black and nonminority scores on high school tests for several years. On the other hand, the SAT is inconsistent with this pattern; the relative gains of black students on that test ended in 1981 and 1982.

26. National Assessment of Educational Progress, *The Reading Report Card*, Figure 3.2.

Despite these changes, the gap between the average scores of black and nonminority students remains striking. On the SAT, for example, the average black student's score in 1975 corresponded roughly to the 11th and 12th percentiles among nonminority students on the mathematics and verbal scales, respectively. In 1984, the average black scores had risen to about the 16th percentile among nonminority scores on both scales.[27] While some other tests show smaller average differences than the SAT, the gap nonetheless remains large by virtually any measure.

Hispanic Students. In national samples, Hispanic students on average show substantially lower levels of achievement than nonminority students, though somewhat higher achievement than blacks. In recent years, the average achievement of Hispanic students, like that of blacks, has risen relative to that of nonminority students.

Generalizations about achievement trends among Hispanic students, however, are subject to important qualifications. First, the relevant data are more limited than in the case of blacks. More important, the term "Hispanic" subsumes many groups differing in culture of origin, lengtl residence in the United States, relatively fluency in and use of English . Spanish, and other factors that presumably affect educational performance. Thus, trends among Hispanic students as a whole provide only suggestions of trends that might be occurring in more specific groups that are often the targets of specific educational programs--such as children with limited proficiency in English, or the children of migrant farm workers.

With those qualifications in mind, the relative improvement of Hispani .' achievement is apparent in the NAEP reading and mathen.atics assessments (see Figure IV-5), in the SAT, in the Texas state-wide assessment of ninth grade students, and in a comparison of nationally representative samples of high school seniors in 1971 and 1979 (the NLS and HSB comparison).[28] This trend appears not to be limited to one Hispanic group. Relative gains appear among both Mexican-American and Puerto Rican students on the SAT and among both Mexican Americans and "other Hispanics" in the NLS and HSB comparison, although the improvement among Mexican Americans is in several instances greater.[29] The annual

[27] These estimates are based on nonminority within-group standard deviations in 1983-1984 reportec n Solomon Arbeiter, *Profiles of College-Bound Seniors, 1984* (New York: The College Entrance Examination Board, 1984), p. 81.

[28] Rock and others, *Factors Associated with Decline of Test Scores*, Appendix D. In this instance, however, the differences in 'he trends shown by Hispanics and nonminority .tudents are slight.

[29] The changes in these Hispanic subgroups in the NLS and HSB comparison, however, appear somewhat unstable and are statistically not significantly different from no change.

SAT data suggest that among Hispanics--as among blacks--the achievement decline ended a few years earlier than it did among nonminority students.

DIFFERENCES IN TRENDS AMONG TYPES OF SCHOOLS AND COMMUNITIES

While the achievement decline was pervasive, it has not been entirely uniform among different types of communities and schools. This section discusses the relative trends in three specific types of schools and communities about which data are available:

- Disadvantaged urban communities;
- Schools with different concentrations of ethnic minorities; and
- Private schools.

Disadvantaged Urban Communities

Since 1970, 9- and 13-year-olds in disadvantaged urban communities gained ground relative to the nation as a whole on the NAEP mathematics and reading assessments (see Tables IV-1 and IV-2). 30/ In contrast, 17-year-olds in disadvantaged urban communities showed no relative gains in mathematics, and their small relative gains in reading occurred entirely between 1979 and 1983. In two instances--in reading at age 9, and in mathematics at age 13--more than a third of the gap between disadvantaged-urban communities and the nation as a whole was overcome since the early 1970s. 31/

30. For a school to be defined as "disadvantaged urban," it had to be located within either the city limits or the urban fringe of a city of at least 200,000 people (or twin or triplet cities with combined populations over 200,000); and it had to serve a community that had unusually few managerial and professional personnel and atypically many unemployed adults and adults on welfare. The latter criterion was implemented through four steps: asking the principal of each school to estimate the proportion of students whose parents fell into those categories; summing the percentages on welfare and unemployed; subtracting the percentage professional or managerial; and selecting the schools that constituted the top 10 percent on the resulting index. (Westat Corporation, unpublished NAEP documentation).

31. In the case of mathematics, however, the amount by which the gap closed can be considered only approximate, for the 1972 average scores are only estimates. See footnote A, Table IV-1.

TABLE IV-1. AVERAGE MATHEMATICS ACHIEVEMENT IN DISADVANTAGED URBAN COMMUNITIES AND IN THE NATION, NAEP, 1972-1981
(Average percent of items correctly answered)

Group	1972 (Estimated) a/	1977	1981	Percent Change 1972-1981
	Age 9			
Nation	56.7	55.4	56.4	-1
Disadvantaged Urban	41.9	44.4	45.5	9
Nation Minus Disadvantaged Urban	14.8	11.0	10.9	-26
	Age 13			
Nation	58.6	56.6	60.5	3
Disadvantaged Urban	41.5	43.5	49.3	19
Nation Minus Disadvantaged Urban	17.1	13.1	11.2	-35
	Age 17			
Nation	64.0	60.4	60.2	-6
Disadvantaged Urban	51.5	45.8	47.7	-7
Nation Minus Disadvantaged Urban	12.5	14.6	12.5	0

SOURCES: CBO calculations based on National Assessment of Educational Progress, *The Third National Mathematics Assessment: Results, Trends, and Issues* (1983), Tables 5.1 and 5.2; and *Mathematics Technical Report: Summary Volume* (1980), Tables 2, 3, and 4.

a. These estimates for 1972 differ from published NAEP results for the 1972 assessment. The published results for that year are based either on the 1972 item pool or on the items used in both 1972 and 1977, while the trend results comparing the 1977 and 1981 assessments reflect items used in both the 1977 and 1981 assessments. In order to circumvent the large disparities in the item sets, 1972 results were estimated here by adjusting the 1977 results (on the items used in 1977 and 1981) by the 1972-to-1977 change (on the items used in 1972 and 1977).

TABLE IV-2. AVERAGE READING ACHIEVEMENT IN DISADVANTAGED URBAN COMMUNITIES AND IN THE NATION, NAEP, 1970-1983
(Average proficiency scores)

Group	1970	1974	1979	1983	Percent Change 1970-1983
		Age 9			
Nation	207	210	214	213	3
Disadvantaged Urban	178	185	186	194	9
Nation Minus Disadvantaged Urban	29	25	28	19	-34
		Age 13			
Nation	254	255	257	258	2
Disadvantaged Urban	232	229	242	240	3
Nation Minus Disadvantaged Urban	22	26	16	18	-18
		Age 17			
Nation	284	285	285	288	1
Disadvantaged Urban	259	261	258	266	3
Nation Minus Disadvantaged Urban	25	24	26	22	-12

SOURCES: National Assessment of Educational Progress, *The Reading Report Card*, Data Appendix.

NOTE: Details might not add to totals because of rounding.

Schools With High or Low Concentrations of Minority Students

Although information on the relative trends in high- and low-minority schools is limited, such data as are available suggest that, relative to the nation as a whole, high-minority schools have gained in achievement while low-minority schools have lost ground. While the available analyses of these data do not clarify whether the gains of minority students have been larger or smaller in high-minority schools, they do indicate that the relative gains of minority students as a group cannot be attributed entirely to improved performance of those attending low-minority schools. At all ages, mathematics gains between the last two National Assessments (1977 and 1981) were several times as large in schools that had minority enrollments of at least 40 percent than in other schools (see Table IV-3). Similarly, in a comparison of the HSB and NLS test results, seniors in low-minority schools --defined as at least 90 percent nonminority--showed, on average, larger declines from 1972 to 1980 than did seniors in other schools. In the case of vocabulary, the decline in low-minority schools was 83 percent larger than in other schools. The difference was about half that size in mathematics, and a fourth in reading. [32]

Private Schools

The achievement decline occurred among high school students in private as well as public schools. Moreover, it appears to have been nearly as large among private school students in reading and vocabulary, although somewhat smaller in mathematics (if tests of reading, vocabulary, and mathematics administered to seniors during the last half of the decline are an adequate indication). [33] Beyond that, very little can be said about the relative trends among private school students, because of the extremely sparse data. For example, whether the upturn in achievement found in public school and nationally representative data--the latter of which is dominated by the far more numerous public school students--occurred in private schools as well is not yet known.

[32] Rock and others, *Factors Associated with Decline of Test Scores*, Appendix D.

[33] Ibid.

TABLE IV-3. AVERAGE MATHEMATICS ACHIEVEMENT IN HIGH-MINORITY AND LOW-MINORITY SCHOOLS, NAEP, 1977 AND 1981 (Average percent of items correctly answered)

Group	1977	1981	Percent Change 1981-1977
	Age 9		
Nation	55.4	56.4	1.8
40 Percent or More Minority	46.4	48.8	5.2
Less than 40 Percent Minority	57.6	58.6	1.7
	Age 13		
Nation	56.6	60.5	6.9
40 Percent or More Minority	45.5	53.6	17.8
Less than 40 Percent Minority	59.6	62.4	4.7
	Age 17		
Nation	60.4	60.2	-0.3
40 Percent or More Minority	47.5	52.3	10.1
Less than 40 Percent Minority	62.4	62.4	0.0

SOURCE: National Assessment of Educational Progress, *The Third National Mathematics Assessment: Results, Trends, and Issues* (1983), Table 5.2.

The SAT decline was found among both private- and public school students.[34] Since the selection changes that contributed to the SAT decline might have been very different among private school students, however, a comparison of the size of the SAT decline in the two groups of students would be risky.

[34] Advisory Panel on the Scholastic Aptitude Test Score Decline, *On Further Examination*, p. 20.

APPENDIXES

APPENDIX A

DESCRIPTION OF MAJOR DATA SOURCES

This Appendix briefly describes the most important data sources used in the text and in other appendixes. These sources are:

- Two college-admissions tests--the Scholastic Aptitude Test (SAT) and the American College Testing Program (ACT) tests;
- The National Assessment of Educational Progress (NAEP);
- The test data from two nationally representative studies of high school students--the National Longitudinal Study of the High School Seniors Class of 1972 (NLS) and the High School and Beyond study (HSB); and
- Annual statewide test data from Iowa.

THE SCHOLASTIC APTITUDE TEST

The Scholastic Aptitude Test (SAT), sponsored by the College Board and administered by the Educational Testing Service, is intended to aid colleges in selecting students for admission. It is perhaps the single best known test in the United States and has figured prominently in discussions of achievement trends for a decade or more.

The SAT is taken by a large number of students, but they constitute a clearly nonrepresentative group. Students taking the test are predominantly those intending to attend college, have higher levels of achievement than does the student body as a whole, and are concentrated in certain geographic regions. In the 1984-1985 school year, the SAT was taken by nearly one million high school students, representing over a third of all graduates and about two-thirds of college-bound graduates. 1/ Nonetheless, it was the

[1] The College Entrance Examination Board, *National College-Bound Seniors, 1985* (New York: The College Board, 1985). The number of high school graduates in the 1984-1985 school year, excluding high school equivalency credentials, has been projected to be about 2.6 million. National Center for Education Statistics, *Projections of Education Statistics to 1990-91* (Washington, D.C.: NCES, 1982), Table 15.

principal college admissions test in only 22 states, which were primarily in the east and on the west coast. 2/

The SAT consists of two tests, one mathematical and one verbal.3/ The verbal test consists of analogies, antonyms, sentence completions, and reading passages. 4/ The mathematics test consists of a variety of problems in arithmetic reasoning, algebra, and geometry that are intended to "require as background mathematics typically taught in grades one through nine" but to "depend less on formal knowledge than on reasoning." 5/

The SAT is designed to predict achievement in college, not to directly assess achievement in secondary schools. Accordingly, the test has been validated primarily by documenting that students scoring higher on the test tend to have higher grades in college. 6/ In contrast, tests intended to assess students' current levels of mastery are typically validated by showing that students scoring higher on the test in question tend to score higher on some other measure of current achievement, such as teachers' evaluations or other achievement tests. 7/

Although the SAT is designed to be a predictor of college performance and was neither intended nor validated as an achievement test, it has often been used as an index of achievement--despite strong objections from the

[2] *State Education Statistics: State Performance Outcomes, Resource Inputs, and Population Characteristics, 1982 and 1984* (Washington, D.C.: U.S. Department of Education, January 1985).

[3] A third scale, the "Test of Standard Written English," was first added on an experimental basis in the mid-1970; it is not discussed in this paper.

[4] The College Board, *College-Bound Seniors.*

[5] Advisory Panel on the Scholastic Aptitude Test Score Decline, *On Further Examination* (New York: The College Board, 1977), p. 9.

[6] Hunter M. Breland, *Population Validity and College Entrance Measures* (New York: The College Board, 1979). It is well established that high SAT scores are associated with higher grades early in college. The extent to which the SAT provides information about likely college performance above and beyond that provided by other indices such as high school grades is a matter of some disagreement. That issue, however, is not germane to the use of SAT scores in this paper. (See, for example, James Crouse, "Does the SAT Help Colleges Make Better Selection Decisions?" *Harvard Educational Review*, vol. 55, May 1985, pp. 195-219; and George H. Hanford, "Yes, the SAT Does Help Colleges," *Harvard Educational Review*, vol. 55, August 1985, pp. 324-331.)

[7] See, for example, Science Research Associates, *SRA Achievement Series, Technical Report #3* (Chicago: SRA, 1981).

College Board. 8/ For example, much of the public debate about declining achievement focused at least in part on the SAT, and the annual compilation of state education statistics by the U.S. Department of Education calls the test a "performance outcome" (rather than a "predictor of performance"). 9/

The SAT is administered several times each year, and the scores obtained in each year are equated, so that any given score should reflect approximately the same level of skill in any year. Annual publications provide detailed tabulations of the scores of the test-taking group as a whole and of a variety of subgroups, such as males, females, and ethnic groups. Data on student characteristics such as these are mostly based on a Student Descriptive Questionnaire (SDQ) completed by students, and the information is therefore subject to distortions stemming from both nonresponse and various kinds of reporting errors.

Data on the SAT extend back longer than those on most other tests, but the long-term data used in this paper are subject to several inconsistencies. Current tabulations by the College Board reflect only the most recent test taken by students who also completed the SDQ--about 90 percent of all SAT candidates. 10/ Average scores from the 1966-1967 through 1970-1971 school years are College Board estimates of the averages that would have been obtained if such tabulations had been made for those years. Data from the 1956 through 1965 school years are based on the average of all scores, which includes multiple scores by those taking the SAT more than once. 11/ The published data on these averages of all scores were adjusted by subtracting from them the slight difference in 1966 between that average and the average based on only the most recent of each individual's scores. Trend data on the proportion of SAT scores above specific thresholds were subject to a similar discontinuity and were similarly adjusted, but in that case the adjustment was based on the average discrepancy in averages over

[8] See, for example, statement by Daniel B. Taylor, Senior Vice President, the College Board, before the House Subcommittee on Elementary, Secondary, and Vocational Education, Committee on Education and Labor, January 31, 1984.

[9] *State Education Statistics: State Performance Outcomes, Resource Inputs, and Population Characteristics, 1982 and 1984* (Washington, D.C.: U.S. Department of Education, January 1985).

[10] The College Board, *College-Bound Seniors, 1985*, p. 4.

[11] Hunter M. Breland, *The SAT Score Decline: A Summary of Related Research* (New York: The College Board, 1976), Table 1.

a four-year period (1971 through 1974) for which both averages were available. 12/

THE AMERICAN COLLEGE TESTING PROGRAM TESTS

The American College Testing Program (ACT) tests, like the SAT, are intended as an aid in selecting students for admission to college. The ACT tests were taken by about 739,000 high-school students in the class of 1984-1985--over a fourth of all graduates. Although the ACT battery is taken by fewer students than is the SAT, it is the predominant college-admissions test in 28 states--primarily in the Midwest, the western mountain states, and parts of the Southeast. 13/

Although also intended to predict success in post-secondary education, the ACT is conceptually distinct from the SAT and is in some senses intended to be more of a test of achievement. The ACT is more "curriculum based" than is the SAT, relying on both reasoning ability and knowledge of subject-matter fields. Despite its intentional reliance on subject-matter knowledge, however, the ACT contains many "analytical, problem-solving exercises and few measures of narrow skills." 14/

The ACT battery consists of subject-matter tests in English, mathematics, social studies, and natural science, yielding four subject-specific scores as well as a composite score. The English test is a test of usage, tapping skills such as grammar, sentence structure, and paragraph organization. The mathematics test is dominated by questions on arithmetic and algebraic reasoning, geometry, and intermediate algebra, but a fourth of the test is devoted to arithmetic and algebraic operations, number concepts, and advanced topics. The social studies test includes aspects of history, government, anthropology, sociology, psychology, and economics. The

[12] Ibid., Table 5.

[13] American College Testing Program, *National ACT Assessment Results, 1984-1985: Executive Summary* (Iowa City: ACT, 1985); U.S. Department of Education, *State Education Statistics*.

[14] *Content of the Tests in the ACT Assessment* (Iowa City: American College Testing Program, undated).

natural sciences test is about evenly divided between chemistry, physics, other physical sciences, and biology. 15/

The ACT is reported and equated annually. Trend data reflecting subgroups of students are available but are less extensive than those available for the SAT.

The long-term ACT trend data used in this paper are subject to one inconsistency. Scores from 1969 on are taken from internally consistent tabulations published by ACT. 16/ Earlier data are adapted from tabulations that differ from the more recent data in including scores from "residual" testing of students on college campuses, who have lower average scores than those taking the test before college. 17/ These earlier averages were adjusted by adding to them the small difference in 1969 between them and the averages consistent with later data. 18/

THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

The National Assessment of Educational Progress (NAEP) is a critical indicator of achievement trends, for it alone among current data sources provides repeated testing of nearly representative samples of the national student population.

Before the NAEP was begun, available data often provided an indication of achievement patterns and trends in smaller areas--that is, in schools, districts, or occasionally states. But variations in assessment methods from one jurisdiction to another precluded using these data as an unambiguous indicator of achievement across the entire nation.

In contrast, the NAEP was designed to be a measure of the performance of the nation's elementary and secondary educational system as a whole. It was not intended to duplicate the assessment mechanisms already in place. For example, it was intended to assess relatively general levels of

[15] American College Testing Program, *Content of the Tests.*

[16] For example, *American College Testing Program, National Trend Data for Students Who Take the ACT Assessment* (Iowa City: ACT, undated).

[17] James Maxey, American College Testing Program, personal communication, April 1984.

[18] The unadjusted earlier data are in L. A. Munday, *Declining Admissions Test Scores* (Iowa City: American College Testing Program, 1976), Table 3.

knowledge, and it was not designed to differentiate among individuals. It was to supplement those other measures by providing a consistent, broad measure of the achievement of a largely representative sample of the nation's youth that would be periodically repeated. 19/

Since 1969, the NAEP has provided periodic testing of 9-, 13-, and 17-year-old students in 10 subject areas. The intervals between assessments in any subject area typically have ranged from three to five years. The best known assessments are in the areas of reading, writing, mathematics, science, and social studies. 20/

Although the NAEP is nearly representative of students nationwide, it excludes several important groups. In most instances, the NAEP has tested only those individuals still in school. 21/ In the case of 17-year-olds, this practice leads to results that are probably quite different from those that would be obtained if all 17-year-olds were tested, since dropouts are numerous in that age group and tend to be low achievers. The overall average score is thus higher than it would be, and comparisons between groups (ethnic groups, regions, and so on) reflect differences in dropout rates as well as achievement differences in the entire age cohort. In addition, handicapped students and those with limited proficiency in English are excluded from testing, although the definition of those categories can vary somewhat from one participating school to another. Both of these exclusions are germane to the assessment of trends, since the period over which the NAEP has been conducted saw the passage of the Education of the Handicapped Act (which most likely increased the number of handicapped students in regular school programs markedly) and rapid immigration from Latin America and Asia. Finally, participating schools have some discretion to exclude other students who cannot be assessed properly. 22/

19. *Director's Report to the Congress on the National Assessment of Educational Progress* (Washington, D.C.: National Institute of Education, December 1982).

20. At various times, the National Assessment has included tests of other groups and subjects that are not considered here.

21. Brief descriptions of the NAEP sampling procedure are provided in a number of publications. See, for example, National Assessment of Educational Progress, *Mathematical Technical Report: Summary Volume* (Denver: NAEP/Education Commission of the States, 1980), Chapter 1.

22. Lawrence Rudner, Office of Educational Research and Improvement, U.S. Department of Education, personal communication, December 1985.

The NAEP tests are designed to assess a range of skills varying in difficulty. In mathematics, for example, the easiest items tap recall of factual information and simple arithmetic computation. More difficult items require an ability to manipulate algebraic expressions, to comprehend and explain mathematical relationships, and to apply skills in solving problems. 23/

For the purposes of this paper, the principal advantages of the NAEP are its nearly representative sampling, its diversity of subject areas and levels of skills, and a considerable amount of background information. A variety of characteristics of students, schools, and communities were ascertained through student, teacher, and school questionnaires. These data permit comparisons of trends, for example, among ethnic groups, geographic regions, and schools with high and low minority enrollments.

These advantages are mitigated, not only by the time intervals between assessments, but also by the forms in which data were presented and the lack of formal equating of scores from one assessment to another. Until recently, scores were generally only reported as the percentage of items answered correctly--a scaling that has some intuitive appeal but one that poses serious problems in gauging trends and, especially, in comparing trends among groups. 24/ In addition, information on the standard deviation of average scores was often not reported or retained, limiting the extent to which the severity of trends could be quantified and compared with that on other tests. Beginning with the most recent assessment of reading, these problems have in large part been solved, but most of the trend data remain in the original form. Scores were also not formally equated until recently, posing problems in the interpretation of trends that were compounded by periodic alteration of the content of the tests. A frequent, but not fully adequate, response to this problem in the published NAEP data was to base comparisons only on items shared by adjacent assessments.

THE NATIONAL LONGITUDINAL SURVEY AND HIGH SCHOOL AND BEYOND

Two nationally representative longitudinal studies of high school students--the National Longitudinal Study of the High School Seniors Class of 1972

23. See, for example, National Assessment of Educational Progress, *Changes in Mathematical Achievement, 1973-78* (Denver: NAEP/Education Commission of the States, 1979).

24. See Chapter II.

(NLS) and the High School and Beyond study (HSB)--provide comparative information on the achievement of seniors in the 1971 and 1979 school years. 25/

Both studies included a variety of cognitive tests, of which three that were administered in both years--vocabulary, reading, and mathematics--can be considered measures of achievement. 26/ The reading and vocabulary tests were identical in the two studies; in mathematics, about half of the items were identical, a fourth were altered in relatively minor respects, and the remainder were new.

In one recent study, the scores on the NLS and HSB tests in those three subject areas were equated, providing an indication of changes in performance over the eight years. 27/ All comparisons of the NLS and HSB in this paper are drawn from that study.

Information is available in the NLS and HSB about a considerable number of important student, school, and community variables, making possible both comparisons of achievement changes in different groups and estimation of the effects of population changes (such as trends in the ethnic composition of the school-age population) on average test scores. This information is derived from school records, school questionnaires, and teacher questionnaires, as well as from student self-reports, which increases the validity of some of the information compared with that obtained solely through student questionnaires. Moreover, in some instances, it permits information from one source to be confirmed by comparing it with that from another.

The usefulness of the NLS and HSB for analyzing achievement trends is limited by several factors, however. The absence of earlier, comparable

25. The NLS and HSB tests were administered in the springs of 1972 and 1980, and most discussions of them refer to those calendar years. In order to be consistent with the treatment of other tests, however, this paper refers instead to the school years in which the tests were administered.

26. Other tests tapped basic cognitive skills but could not be considered measures of achievement. For example, a mosaic comparisons test was included in 1972 as an index of "perceptual speed and accuracy." For a brief description of the two test batteries, see Donald A. Rock, Ruth B. Ekstrom, Margaret E. Goertz, Thomas L. Hilton, and Judith Pollack, *Factors Associated with Decline of Test Scores of High School Seniors, 1972 to 1980* (Washington, D.C.: Center for Statistics, U.S. Department of Education, 1985) Chapter II.

27. Donald Rock and others, *Factors Associated with Decline of Test Scores.*

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).

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 otner 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.

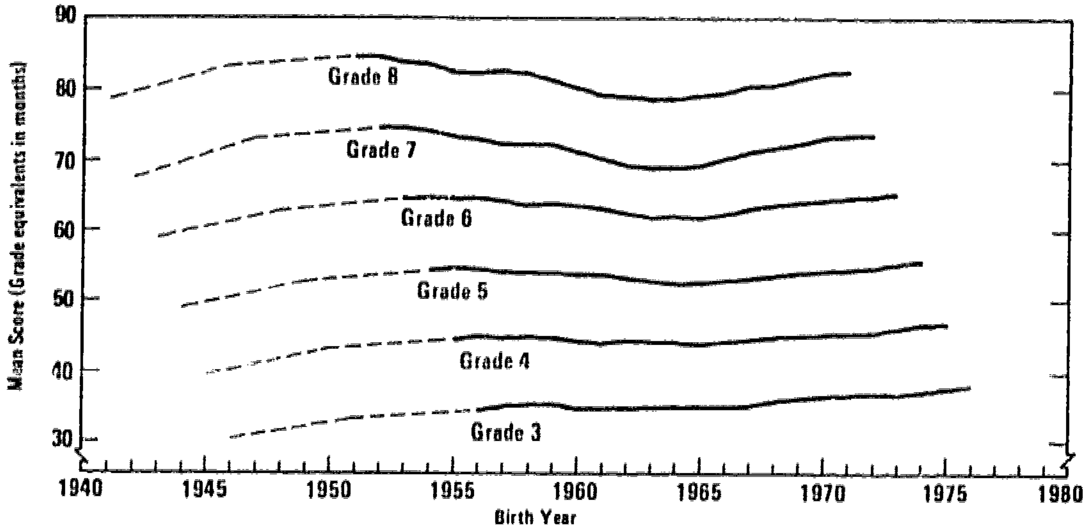
The greatest weakness of the ITBS-IA data is the fact that Iowa is in several important respects atypical of the United States. By some measures, average achievement in the elementary and secondary grades is nearly a grade higher in Iowa than in the nation as a whole. [4] In addition, the student body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



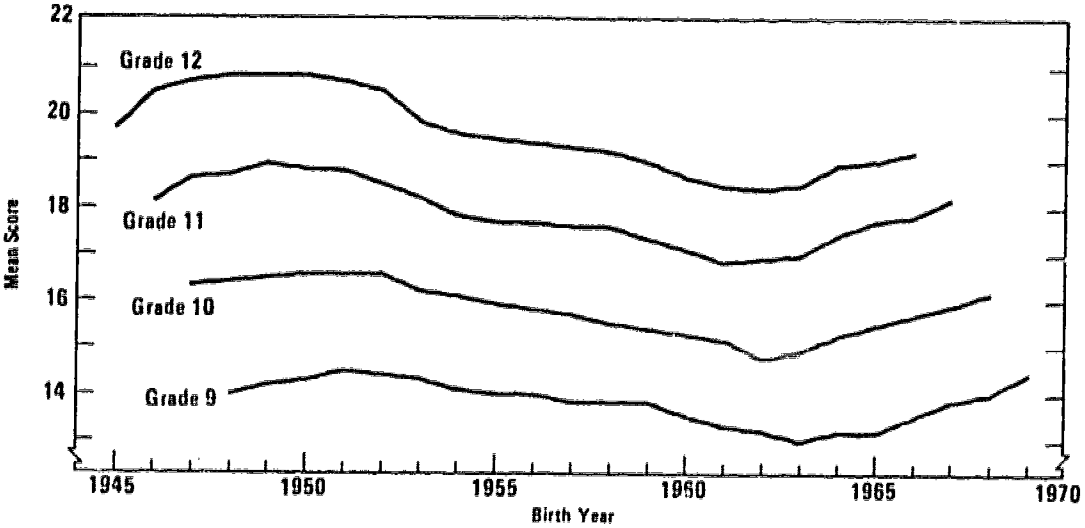
SOURCE: CBO calculations based on "Iowa Basic Skills Testing Program, Achievement Trends in Iowa: 1955-1985" (Iowa Testing Programs, unpublished and undated material).

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).

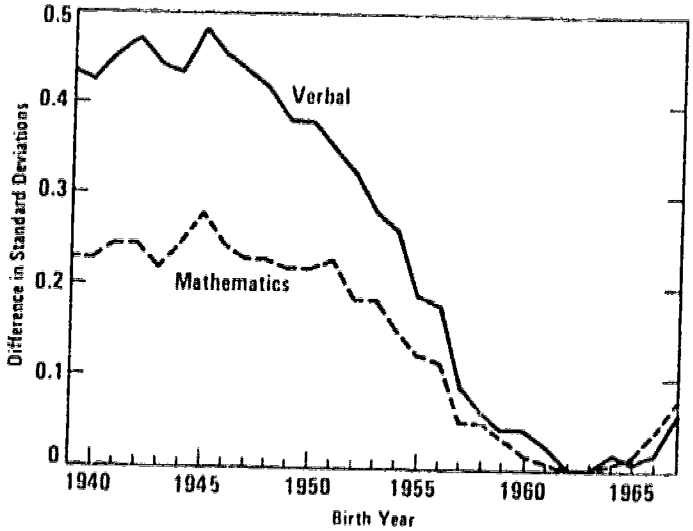
Enough is known about self-selection of students taking the SAT to know that those taking it are not representative of high school seniors in general. Not enough is known, however, to control fully for the nonrepresentativeness of the SAT sample. On the other hand, while compositional changes--that is, 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)



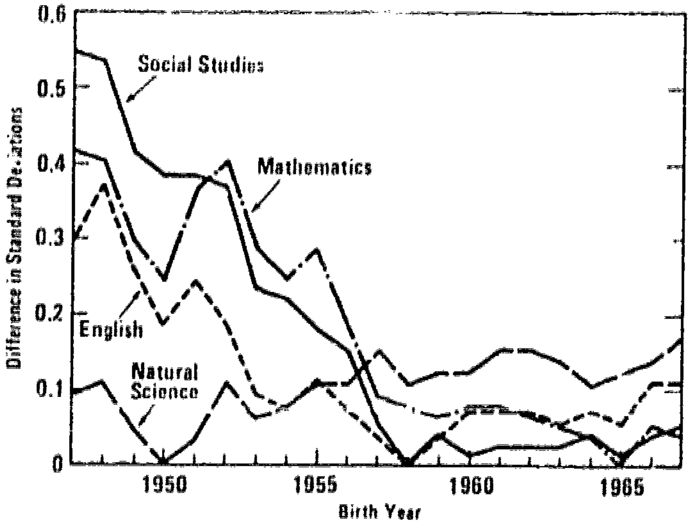
SOURCES: CBO calculations based on Hunter M. Breland, *The SAT Score Decline: A Summary of Related Research* (New York: The College Board, 1987); and The College Entrance Examination Board, *National College-Bound Seniors, 1985* (New York: The College Board, 1985).

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/

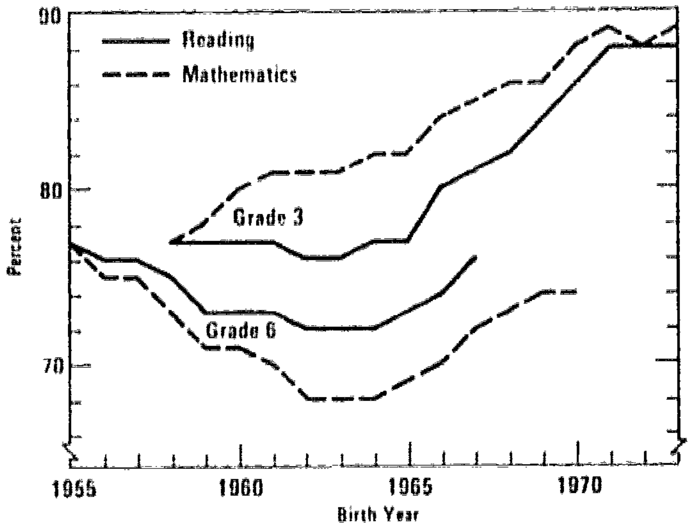
Three of the four tests--reading at both grade levels, and mathematics at grade six--conform to the cohort model suggested by the ITBS, the ITED, and the SAT. These three tests stopped declining with the birth cohort of 1962 and began improving markedly within a few years (see Figure B-5). Because the 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.

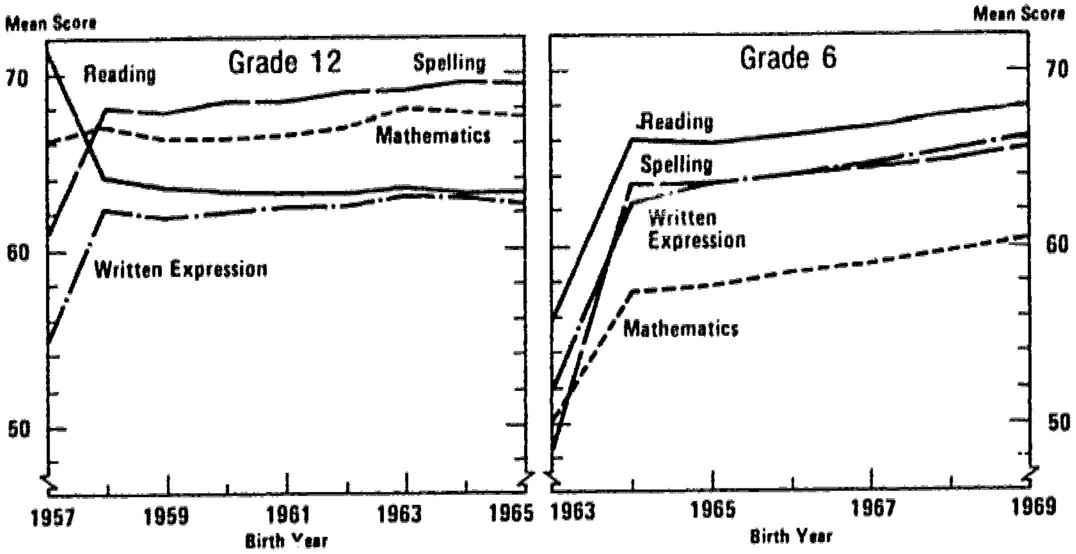
7. Division of Educational Testing, *Student Achievement in New York State 1982-83* (Albany: New York State Education Department, January 1984).

Figure B-5.
Percent of New York Students Scoring Above Refeience 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.

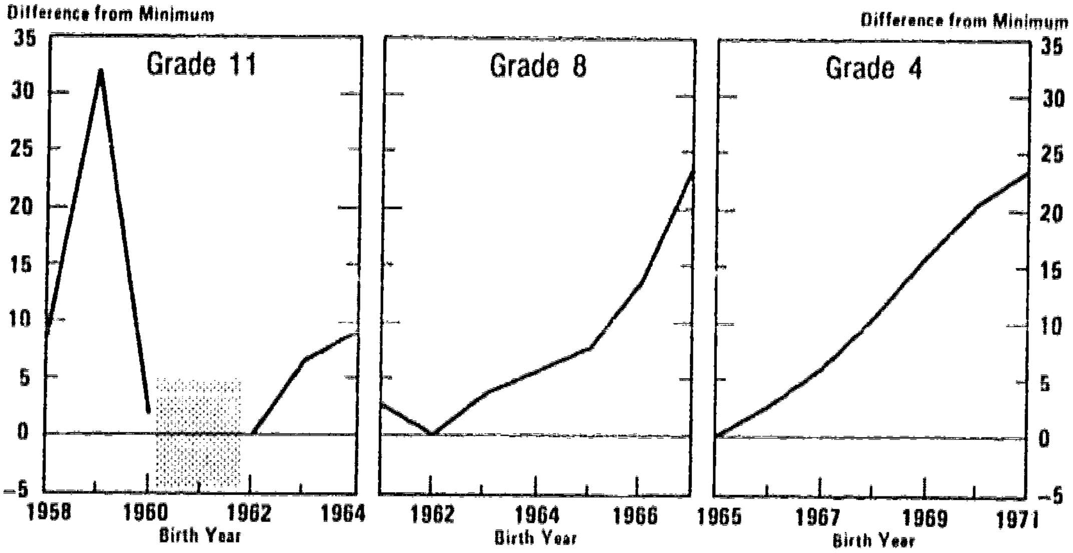
The Virginia assessment data provide a weaker test of the cohort and period models than do the data series above, but they provide a stronger test than do some of the intermittent data series discussed below. The relevant fourth-grade data begin only with the 1965 birth cohort, which is too recent to 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)

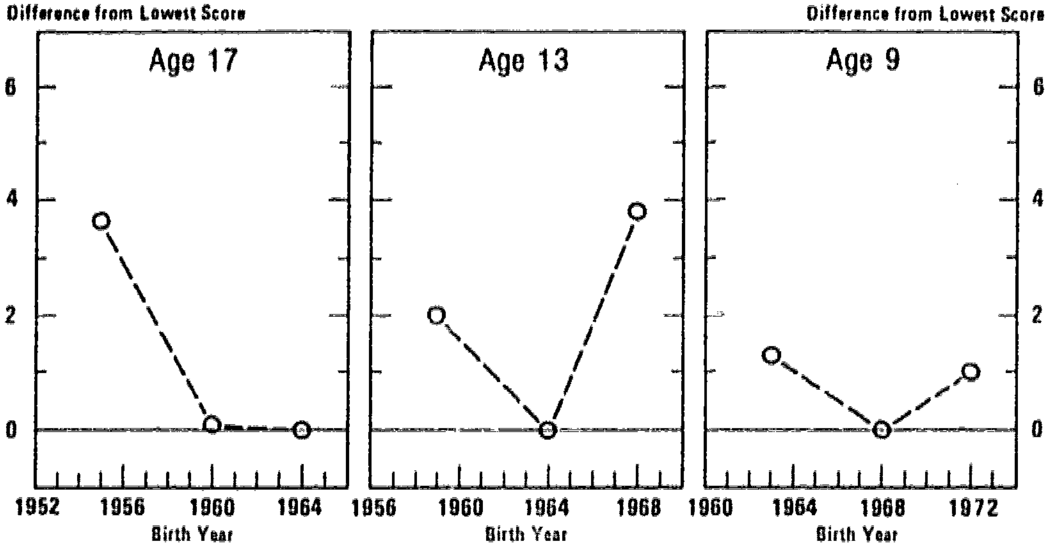


SOURCE: CBO calculations based on S. John Davis and R. L. Boyer, *Memorandum to Division Superintendents: State Testing Program Results, 1980 1981* (Ricl mond: Commonwealth of Virginia Department of Education, 1981).

there was only one recent change in the direction of the trend--that is, only one recent period each of dec'ine 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

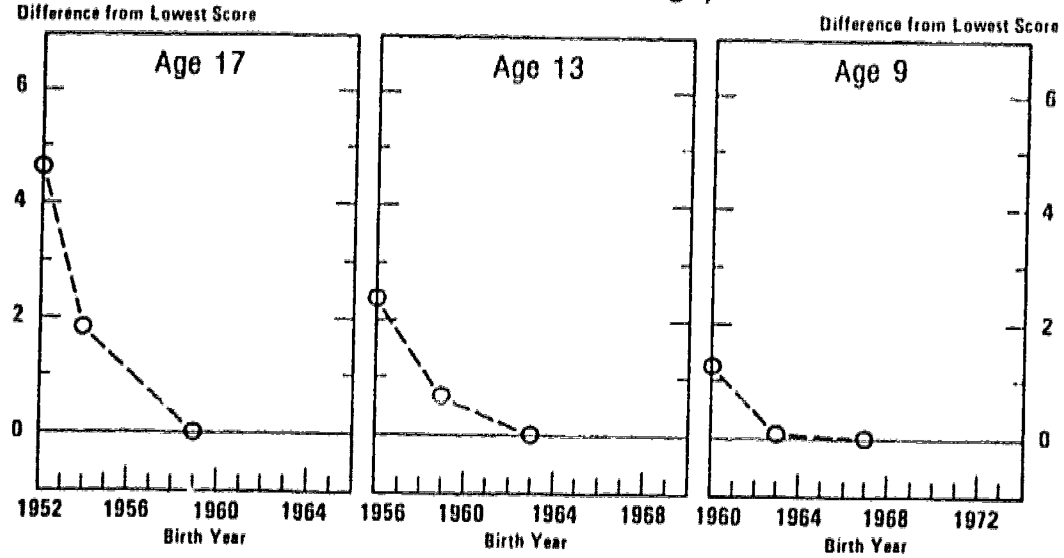
The ITBS, like most commercial standardized elementary and secondary tests, is renormed approximately once every seven years. The ITBS norming data reported here, unlike the ITBS-IA data described above, is based on national samples of students. [8]

[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.

Figure B-9.

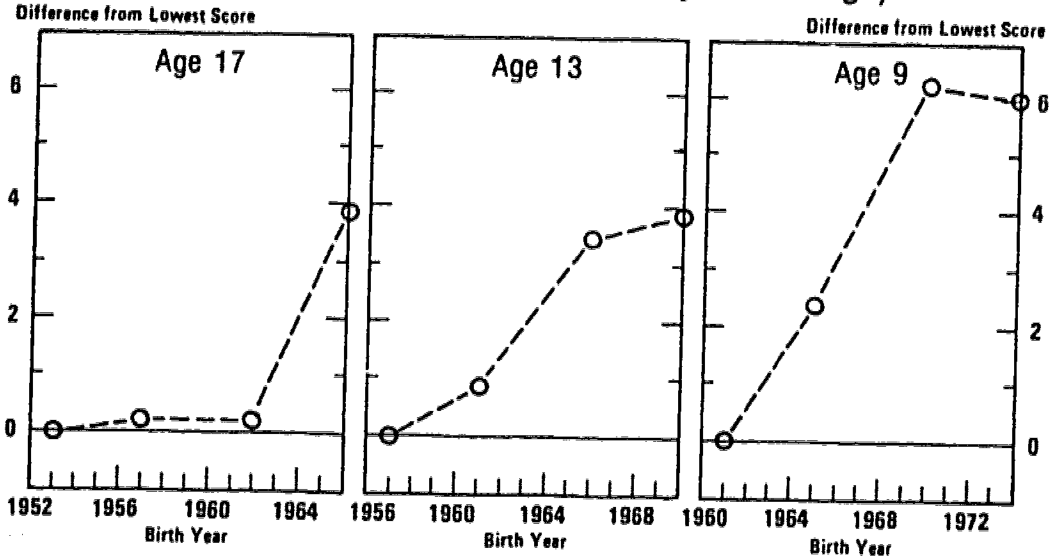
NAEP Science Scores (By birth year and age)



SOURCE: CBO calculations based on National Assessment of Educational Progress, *Three National Assessments of Science: Changes in Achievement, 1969-77* (Denver: NAEP/Education Commission of the States, 1978).

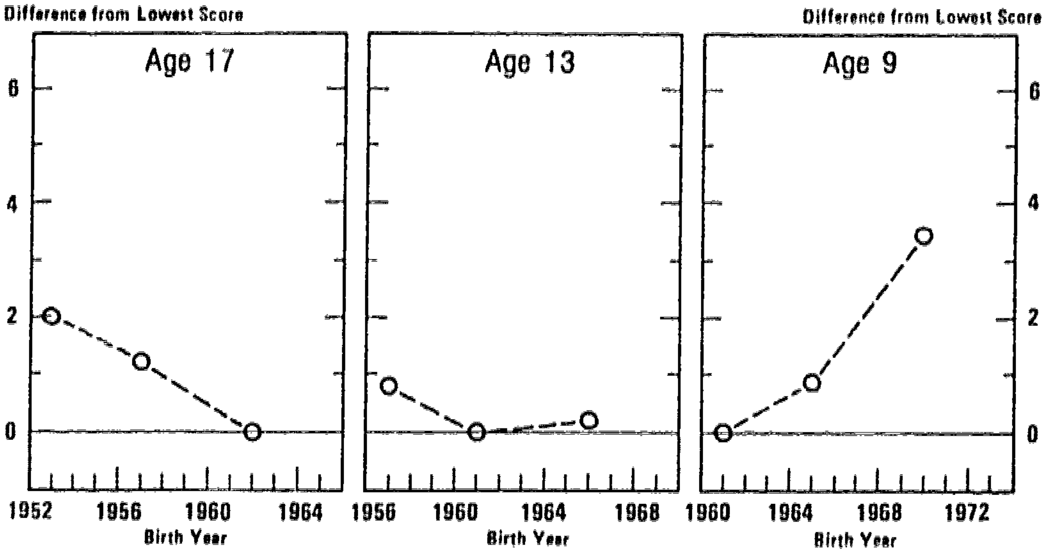
Figure B-10.

NAEP Reading Proficiency Scores (By birth year and age)



SOURCE: CBO calculations based on National Assessment of Educational Progress, *The Reading Report Card: Progress Toward Excellence in Our Schools* (Princeton: NAEP/Educational Testing Service, 1985).

Figure B-11.
NAEP Reading (Inferential Comprehension) Scores
(By birth year and age)



SOURCE: CBO calculations based on National Assessment of Educational Progress, *Three National Assessments of Reading: Changes in Performance, 1970-80* (Der ver: NAEP/Education Commission of the States, 1981).

Norming data offer even weaker evidence than does the NAEP for testing the cohort and period models. Like the NAEP and all other intermittent data, norming data cannot precisely pinpoint the timing of a turnaround in achievemen.t trends. In addition, norming data usually have even longer gaps between test years than those in the National Assessment (most often, seven years). These long gaps further exacerbate the uncertainty. Norming data generally also entail testing all grade levels in all subjects at the same time. In conjunction with the long period between renorming, this factor can force the trend data to appear to be a period effect even if the true underlying pattern is a cohort effect. 9/

9. The extent of this bias depends on the time span between normings, the range of grades tested, the number of years between a given norming and the true minimum in the trend data, and the slope of the curves on both sides of the minimum. For example, suppose that grades four through six are tested in 1972 and 1979 and that the true trend is a cohort model, with grade four reaching its low point in 1972, grade five in 1973, and so on. If the declines and upturns in each grade are reasonably similar in severity, all three grades will show their lowest scores in the 1972 norming sample. If, however, the testing continues through grade 12, the older grades--beginning with grade eight or nine--would probably show their lowest scores in the 1979 norming sample.

Taken together, the ITBS norming data can be seen as consistent with either a period or a cohort model. But they do suggest--albeit weakly--that if the cohort model is correct, the low point might be a few cohorts later than in the Iowa, SAT, and New York data. In all grades from fourth through eighth, the scores of students in the norming sample reached their lowest observed levels with the norming of 1977-1978, corresponding to the birth cohorts of 1964 through 1968 (see Figure B-12).[10] If the decline reached its end with the birth cohort of 1963, for example, one might expect fourth-and fifth-grade scores to be lowest in the prior (1970-71) norming.

The California Test of Basic Skills (CTBS) Norming Data

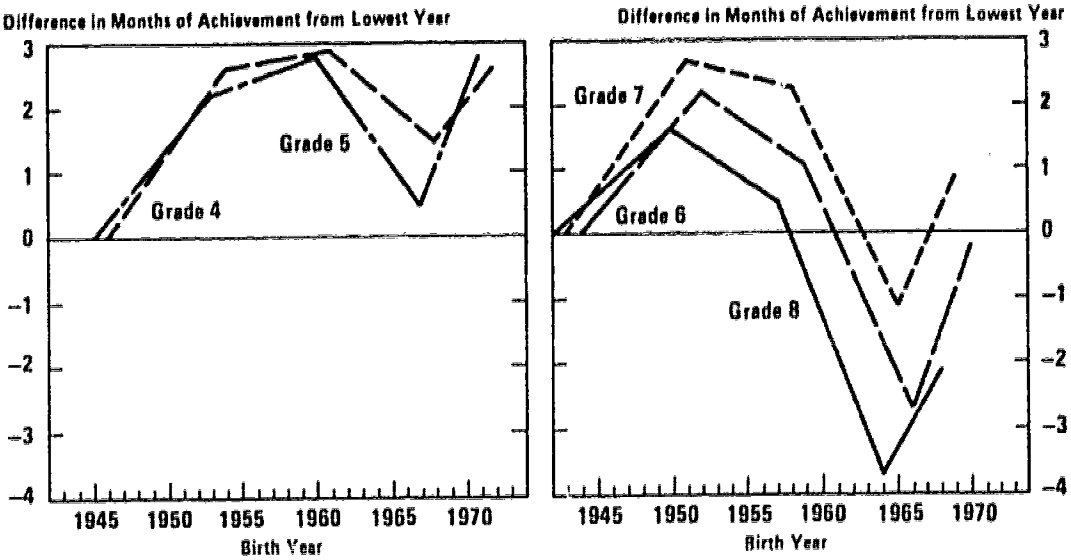
An equating study of the most recent (1973 and 1980) normings of the CTBS provides a somewhat stronger test of the two models, for the large span of grades tested (first through twelfth) in part compensates for the long interval between the two test dates.

If the cohort model and the timing suggested by the Iowa, SAT, and New York data are correct, the CTBS data should show increases that are sizable in the elementary grades, gradually decrease in size in the junior-high grades, and are replaced by declines in the senior-high grades. In grade five and below, both norming samples comprise cohorts born in 1963 or later--that is, cohorts that produced increasing scores in the other data bases. In grades 6 through 11, the norming samples comprise varying mixes of post-1963 and pre-1963 birth cohorts, and the increases among the former should tend to offset the declines among the latter. Finally, both grade-12 samples were born in 1963 or earlier, so if the decline ended in 1963, the change at that grade level would reflect only years of declining achievement.

The changes in the CTBS norming samples largely conform to these predictions from the cohort model. With one exception, all comparisons at grade nine and below showed increases from 1973 to 1980, with a tendency for the largest gains to be in the lowest grades. For example, in the fall testing, the achievement of a third-grade student scoring at the 34th percentile in 1980 corresponded roughly to that of the median student in 1973, while in grade eight, a student would have had to reach the 46th percentile to score at the level of the median student of seven years earlier.

[10] In grade three, average scores increased with every norming sample after the initial (1955) one--mirroring the negligible decline in third-grade scores in the ITBS-IA data --and are excluded from this discussion.

Figure B-12.
ITBS National Norming Data (By birth year and grade)



SOURCES: CBO calculations based on A. N. Hieronymus, E. F. Lindquist, and H. D. Hoover, *Iowa Test of Basic Skills: Manual For School Administrators* (Chicago: Riverside, 1982); and *The Development of the 1982 Norms for the Iowa Tests of Basic Skills* (Chicago; Riverside, 1983).

In contrast, students in the eleventh and twelfth grades showed a drop in achievement during that period. 11/

The California Achievement Tests (CAT) Norming Data

The 1970 and 1977 normings of the CAT were equated to each other and can be used in the same way as the 1973 and 1980 CTBS to test the cohort and period models. The two editions of the CAT, however, were more dissimilar from one another, making the procedure riskier.

Because the CAT was renormed three years earlier than the CTBS, one would expect the observed changes to switch from increases to decreases

11. California Test Bureau, McGraw Hill, unpublished tabulations. The most salient exception to this pattern occurred among ninth-grade students, who showed larger gains than any students above grade three.

three grades younger. Specifically, if the decline followed the cohort model and reached its low point with the 1963 birth cohort, one would expect grades nine and above to reflect only years of decline, while grades three through eight would reflect varying mixes of increasing and decreasing years. Only grades one and two would reflect solely increasing years, and those are grades in which the decline appears never to have occurred.

The results of the CAT renorming study largely conform to these predictions based on the cohort model. Grades one and two showed gains of over 0.6 standard deviation. These increases rapidly tapered off with increasing age, so that grades five and six showed essentially no change. Grades seven and eight showed declines of less than 0.2 standard deviation, while the higher grades all showed drops larger than 0.3 standard deviation. 12/

AN AGGREGATE TEST OF THE COHORT AND PERIOD MODELS

Another method of testing the cohort and period models is to assess which model yields the least variable estimates of the timing of the end of the decline, considering only those continuous data bases that show a clear low point. That is, the timing of the decline's end can be estimated in terms of both test years and birth cohorts, and the relative variation in those estimates indicates which of the models fits the data more closely. This approach, however, suffers from the relatively small number of data bases that can be applied.

Among the data bases meeting these criteria, the cohort model fits the data more closely than does the period model (Table B-1). The end of the decline, expressed in test years, showed a mean of 1976 and a 12-year range (from 1970 to 1982). When expressed in terms of birth cohorts, the decline's end showed a mean of 1962 and a range of only seven years (from 1958 to 1965). The standard deviation of the estimate is roughly 60 percent larger when test years are used. 13/

12. California Test Bureau, McGraw Hill, unpublished tabulations.

13. There are several ambiguities, noted in the text above, in specifying single years as the end of the decline in each data series, and these uncertainties apply to the patterns shown in Table B-1 as well. The most striking ambiguity entails the ACT mathematics assessment, which continued to decline, though slightly and inconsistently, for several years after the substantial decline ended. Table B-1 uses the year that the decline ended entirely. Substituting the year that the major decline in mathematics scores ended (1976), however, would not alter the conclusions. While it would make the relative fit of the cohort and period models more similar, the cohort model would still fit appreciably better.

The closer fit of the cohort model is much more striking if the ACT is excluded. The ACT is anomalous in two respects among continuous data bases showing an achievement decline--the early end of its decline and the lack of a subsequent upturn. Because these anomalies are unexplained, retesting the cohort model without the ACT seems warranted. When the ACT is excluded, the test years marking the end of the decline shows a nine-year range (from 1970 to 1979), while the birth cohorts show only a three-year range (from 1961 to 1964). Similarly, the difference between the test-year and cohort-year standard deviations is much larger--the former is nearly 3.5 times as large as the latter.

TABLE B-1. TIMING OF THE END OF THE ACHIEVEMENT DECLINE, BY TEST (Test years and birth years of group showing lowest score) a/

Test	Grade	Test Year	Birth Year
ACT Mathematics	12	1982	1965
ACT English	12	1975	1958
ACT Social Studies	12	1975	1958
SAT Verbal	12	1979	1962
SAT Mathematics	12	1979	1962
ITED Iowa Comprehensive	12	1979	1962
ITED Iowa Comprehensive	11	1977	1961
ITED Iowa Comprehensive	10	1977	1962
ITED Iowa Comprehensive	9	1977	1963
ITBS Iowa Comprehensive	8	1976	1963
ITBS Iowa Comprehensive	7	1975	1963
ITBS Iowa Comprehensive	6	1974	1963
ITBS Iowa Comprehensive	5	1974	1964
Virginia Comprehensive	8	1975	1962
Virginia Comprehensive b/	11	1977.5	1961.5
New York Reference-Point Mathematics	6	1973	1962
New York Reference-Point Reading	6	1973	1962
New York Reference-Point Reading	3	1970	1962

Variability of Estimates

Including the ACT

		Test Year	Birth Year
Mean		1976	1962
Standard Deviation (in years)		2.7	1.7
Minimum		1970	1958
Maximum		1982	1965
Range (in years)		12	7

(Continued)

TABLE B-1. (Continued)

Test	Grade	Test Year	Birth Year
Excluding the ACT			
Mean		1976	1962
Standard Deviation (in years)		2.5	0.73
Minimum		1970	1961
Maximum		1979	1964
Range (in years)		9	3

SOURCES: CBO calculations based on American College Testing Program, *National Trend Data for Students Who Take the ACT Assessment* (Iowa City: ACT, undated); The College Entrance Examination Board, *National College-Bound Seniors, 1985* (New York: The College Board, 1985); "Mean ITED Test Scores by Grade and Subtest for the State of Iowa" (Iowa Testing Programs, unpublished and undated tabulations); "Iowa Basic Skills Testing Program, Achievement Trends in Iowa: 1955-1985" (Iowa Testing Programs, unpublished and undated material); S. John Davis and R. L. Boyer, *Memorandum to Division Superintendents: State Testing Program Results, 1980-81* (Richmond: Commonwealth of Virginia Department of Education, 1981); Division of Educational Testing, *Percent of Pupils Scoring Below State Reference Point on Pupil Evaluation Program Tests* (Albany: New York State Education Department, undated).

a. End is last year before increase or stability. See text for explanation of ambiguities involved in specifying one year as the low point in these series.

b. The low point could be either the 1977 or 1978 test years; no data are available for 1977.

APPENDIX C

DIFFERENCES IN TRENDS BY SUBJECT AREA

As discussed in Chapter III, among all of the tests considered in this paper, no single subject area consistently showed the most severe decline in average scores. Nor was the decline consistently more substantial in either "directly" or "indirectly" taught subjects. This appendix provides the information on which those conclusions are based.

Not all of the data sources discussed in this paper could be used for making comparisons among subject areas. Only those tests that included more than one subject area and that could be converted to standard deviations (SDs) could be used, since only in those instances could the relative size of the decline among subject areas be ascertained. The most serious omission for this reason is the National Assessment of Educational Progress; the NAEP staff did not retain sufficient information on SDs to convert published raw scores.

This appendix includes data from tests administered both annually and less frequently, but comparisons among subject areas often have a somewhat different meaning in the two cases. When annual data are available, the beginning and end of the decline in each subject can be ascertained, and the tabulations in this appendix represent the total amount of each decline, regardless of its duration. In those instances, the largest decline need not be the most rapid. A subject showing a slower decline than others, for example, can drop more in total if its decline is sufficiently long in duration.

In the case of tests administered less often than annually, however, the beginning and end of the decline cannot be pinpointed. In those instances, the tabulations in this appendix represent the amount scores dropped during a fixed period for all subjects in one test battery--for example, the period between two normings, or between the National Longitudinal Study (1971) and the High School and Beyond study (1979).[1] If the period used does not include years of rising scores, these comparisons indicate the relative rate of decline among subject areas, as well as the

[1] The NLS and the HSB tests were administered in the springs of 1972 and 1980, respectively--that is, in the 1971 and 1979 school years.

amount of the decrease over that period. The comparisons need not, however, indicate the relative total decline among different subjects, since they cannot take into account differences in the duration of the decline. Moreover, because the time span used can encompass varying periods of rising scores, these comparisons are less reliable than those based on annual data. 2/

The majority of the tests considered here showed the largest declines on language-related subtests, but the exceptions were frequent enough to suggest that this ranking is more a reflection of the attributes of individual tests than an underlying consistency in the achievement trends (see Table C-1). In addition to the SAT, test batteries that showed the greatest decline in language-related tests include the NLS and HSB comparison, the grade 12 Iowa state data (ITED-Iowa), the Illinois Decade Study, and, for the most part, the Project TALENT 15-year comparison (1960 and 1975). In contrast, Iowa state elementary school data (ITBS-Iowa) show the opposite pattern: the decline in mathematics was much more severe than that in any of the language-related subjects. Senior high school norming data for the California Achievement Test (CAT-US) also show a greater decline in mathematics than in other areas. Other test batteries--such as the national norming data for the elementary-level Iowa test battery (ITBS-US)--show a more complex pattern, with the various language-related tests bracketing the mathematics test in terms of the magnitude of the decline. The ACT showed a slightly larger decline in English than in mathematics. It also showed its largest decline in social studies, however, and no decline at all in science.

The various tests are also inconsistent in terms of the relative declines in "directly taught" and "indirectly taught" subjects. Some of the language-related tests that showed particularly steep declines--such as the vocabulary tests in the Project TALENT data and the NLS-to-HSB comparison--might be viewed as being largely indirectly taught subjects. Other language-related tests that declined markedly, however, presumably are much more reliant on formal instruction--such as the language test in the national ITBS data and the expression test in the national ITED data, both of which are tests of language usage. In addition, mathematics, which has been used as an example of a directly taught subject, showed the steepest decline in several test batteries.

[2] In the case of tests administered less often than annually, the tabulations used here are based on a single interval during which all subjects evidenced declines. If an adjacent interval showed declines in some subjects but not others--as was the case, for example, with the grade-eight ITBS norming data--that adjacent period was ignored.

TABLE C-1. MAGNITUDE OF THE ACHIEVEMENT DECLINE, BY SUBJECT

Test	Grade	Subject	Total Decline (Standard Deviations)
SAT	12	Verbal	0.48
	12	Mathematics	0.28
NLS to HSB	12	Vocabulary	0.22
	12	Reading	0.21
	12	Mathematics	0.14
ITED-US	12	Expression	0.28
	12	Mathematics	0.26
	12	Vocabulary	0.23
ITED-US	10	Mathematics	0.32
	10	Expression	0.29
	10	Vocabulary	0.22
ITED-Iowa	12	Reading a/	0.40
	12	Social Studies	0.36
	12	Expression	0.32
	12	Vocabulary	0.30
	12	Science	0.28
	12	Mathematics	0.27
ITED-Iowa	10	Reading a/	0.32
	10	Mathematics	0.31
	10	Expression	0.29
	10	Social Studies	0.27
	10	Vocabulary	0.25
	10	Science	0.25
ITBS-Iowa	8	Mathematics	0.47
	8	Language	0.37
	8	Reading	0.35
	8	Vocabulary	0.26

(Continued)

TABLE C-1. (Continued)

Test	Grade	Subject	Total Decline (Standard Deviations)
ITBS-Iowa	6	Mathematics	0.38
	6	Language	0.25
	6	Reading	0.17
	6	Vocabulary	0.10
ITBS-US	8	Language	0.32
	8	Mathematics	0.28
	8	Vocabulary	0.23
	8	Reading	0.20
ITBS-US	6	Language	0.32
	6	Mathematics	0.28
	6	Vocabulary	0.19
	6	Reading	0.17
CAT-US	12	Mathematics	0.34
	12	Reading Comprehension	0.24
	12	Vocabulary	0.23
	12	Language	0.18
CAT-US	9	Mathematics	0.30
	9	Language	0.28
	9	Vocabulary	0.21
	9	Reading Comprehension	0.05
ACT	12	Social Studies	0.55
	12	Mathematics	0.42
	12	English	0.37
	12	Science	-0.06
Illinois Decade	11	English 2	0.49
	11	English 1	0.38
	11	Social Studies	0.35
	11	Math 2	0.26
	11	Science	0.19
	11	Math 1	0.05

(Continued)

TABLE C-1. (Continued)

Test	Grade	Subject	Total Decline (Standard Deviations)
Talent 15-Year Follow-Up	9,10,11	Vocabulary	0.40
	9,10,11	English	0.30
	9,10,11	Quantitative Reasoning	0.22
	9,10,11	Reading Comprehension	0.06
	9,10,11	Computation	0.23
	9,10,11	Mathematics	-0.07
	9,10,11	Abstract Reasoning	-0.24
	9,10,11	Creativity	-0.34

SOURCES: CBO calculations based on Hunter M. Breland, *The SAT Score Decline: A Summary of Related Research* (New York: The College Board, 1976); The College Entrance Examination Board, *National College-Bound Seniors, 1978* and *1985* (New York: The College Board, 1985); Donald A. Rock, Ruth B. Ekstrom, Margaret E. Goertz, Thomas L. Hilton, and Judith Pollack, *Factors Associated with Decline of Test Scores of High School Seniors, 1972 to 1980* (Washington: Center for Statistics, U.S. Department of Education, 1985); Robert Forsyth, Iowa Testing Programs, personal communications, April, 1984; "Mean ITED Test Scores by Grade and Subtest for the State of Iowa" (Iowa City: Iowa Testing Programs, undated and unpublished tabulations); "Iowa Basic Skills Testing Program, Achievement Trends in Iowa: 1955-1985" (Iowa City: Iowa Testing Programs, undated and unpublished tabulations); A. N. Hieronymus, E. F. Lindquist, and H. D. Hoover, *Iowa Tests of Basic Skills: Manual For School Administrators* (Chicago: Riverside, 1982); *The Development of the 1982 Norms for the Iowa Tests of Basic Skills* (Chicago: Riverside, 1983); CTB/McGraw-Hill, unpublished tabulations, December 1977; L. A. Munday, *Declining Admissions Test Scores* (Iowa City: American College Testing Program, 1976); American College Testing Program, *National Trend Data for Students Who Take the Act Assessment* (Iowa City: ACT, undated); *Student Achievement in Illinois, 1970 and 1981* (Springfield: Illinois State Board of Eduction, 1983); John C. Flanagan, "Analyzing Changes in School Levels of Achievement Using Project TALENT Ten- and Fifteen-Year Retests," in G. R. Austin and H. Garber (eds.), *The Rise and Fall of National Test Scores* (New York: Academic Press, 1982), pp. 35-49.

NOTE: This table is limited to data that span a sizable portion of the decline and that permit exclusion of the subsequent upturn. Only selected grade levels are presented for the sake of simplicity.

a. This is the "Interpretation of Literary Materials" test. Reading skills are also measured by the ITED social studies and science tests.

APPENDIX D

VARIATION AMONG ACHIEVEMENT SUBGROUPS

As discussed in Chapter IV, there is inconsistent evidence about the relative trends in test scores among different achievement subgroups--that is, among groups of students categorized by their differing levels of achievement. Because this issue has received considerable public attention, and because the conclusions presented in the paper are not entirely in keeping with those presented by some other writers, this appendix provides additional detail about the evidence that underlies the following five generalizations, presented in Chapter IV:

- The achievement decline and the subsequent upturn occurred among both low- and high-achieving students.

- During the mid- and late 1970s--that is, during the end of the achievement decline and the beginning of the subsequent upturn--students in the top achievement quartile on the National Assessment of Educational Progress (the top fourth of all students, when ranked by achievement) lost ground relative to those in the bottom quartile.

- Other data, however, do not consistently suggest a narrowing gap between the top and bottom achievement quartiles. The narrowing evident in the NAEP data might be limited to the short time period of that particular assessment (roughly half of the 1970s), or it might be limited to certain types of tests. Alternatively, more detailed analyses than those now available might show the narrowing to be a more general pattern.

- Test scores of students taking college-admissions tests--currently, about half of all high-school graduates--declined more than those of high school seniors in general, but this difference primarily reflects the changing composition of the group taking those tests rather than a greater decline in achievement among high-achieving students.

- Select students--those scoring highest on tests, taking the most advanced courses, and so on--experienced both the decline and

> the subsequent upturn in achievement. Select students did not show a consistently greater decline than the average student. Indeed, by some measures, select students appear to have gained relative to the average, particularly in the area of mathematics. The sketchiness and inconsistency of data on select students, however, cloud these conclusions.

As noted in Chapter IV, however, both differences and similarities among trends in achievement subgroups must often be taken with a grain of salt. They can be simple artifacts of technical aspects of the tests used--specifically, the scaling of the test, its content, and the measure of change that is reported. For example, if both the top and bottom achievement quartiles show a decline of 5 percentage points in the average number of test items answered correctly, these seemingly equivalent changes could in fact reflect very different real changes in skills. The change would be proportionately larger in the bottom quartile. Moreover, the typical students in each quartile answer very different questions correctly, and only detailed information about the content and difficulty of the additional items answered incorrectly by each quartile would indicate whether the loss of skills in each group are qualitatively or quantitatively similar.[1] Technical solutions of this ambiguity are complex and have rarely been applied to the specific question of relative trends among different achievement subgroups.

The test results cited in this section differ in the certainty of their conclusions about achievement subgroups. At one extreme, the results of the Illinois Decade study are very ambiguous, because two available measures of change lead to different conclusions about achievement subgroup differences. At the other extreme, some--but not all--of the relevant tabulations from the National Assessment are clear-cut, because some show increases in the lowest quartile concurrently with decreases in the top quartile. Use of different scaling or reporting conventions would generally not alter the conclusion of a narrowing achievement gap in those cases.

[1] This ambiguity also arises with other common measures of change, such as scaled-score or standardized-score changes.

Technically, the problem has several aspects. One is that the metrics commonly used are not ratio scales; indeed, they are arguably not even interval scales. The construction of the tests poses additional problems, for a single test is unlikely to be a comparably comprehensive measure of mastery at two very different levels of achievement and therefore may understate the relative change of students at one level. The tabulation and reporting of results further complicates comparisons, since information on the additional items correctly or incorrectly answered is rarely reported, particularly for achievement subgroups.

TRENDS IN THE LOWEST AND HIGHEST QUARTILES

The most extensive and best-known information on the relative trends among students in the top and bottom achievement quartiles is from the NAEP. Relevant information is also available, however, from the SAT, the ACT, the ITBS, and the Illinois Decade Study.

The National Assessment of Educational Progress

In general, the currently available NAEP tabulations show a narrowing of the gap between the top and bottom quartiles in all three age groups (9, 13, and 17) and subjects (reading, science, and mathematics) for which the analysis was conducted. The comparative data, however, span only four or five years during the 1970s. Comparable tabulations of the NAEP are unavailable for the remaining middle half of the student population.

These particular NAEP trends show great variation--changes ranged from sizable improvements to large declines--which complicates comparison of achievement subgroups. This variation probably results in part from the period over which changes were measured--beginning between 1972 and 1974 and ending between 1976 and 1979, depending on the subject tested. Given the cohort pattern shown by the end of the decline, it is likely that these particular assessments of trends among nine-year-olds began about the time that their brief and small decline ended. The trend for 13-year-olds probably spanned the last years of the decline and the first years of the upturn, while the trend among 17-year-olds corresponds roughly to the last years of the decline. Consistent with this cohort pattern, the NAEP data described here show few declines among 9-year-olds, few gains among 17-year-olds, and a more mixed pattern among 13-year-olds. Comparisons are thus clearest if made within any one age group.

In the lowest quartile, nine-year-olds showed improvement in two of three subject areas and no change in the other. This held true for both black and white students (see Table D-1). In the top quartile, black students also showed improvement. White students did not, however; they showed sizable declines in two subjects and no change in a third.

TABLE D-1. RECENT TRENDS IN THE NATIONAL ASSESSMENT, BY ACHIEVEMENT SUBGROUPS AND ETHNICITY

Group	Subject Area		
	Reading	Science	Mathematics
<u>9-Year-Olds in the 4th Grade</u>			
Lowest Quartile			
Black students	Improvement: gain of 8.4 percentage points	No significant change in performance	Improvement: gain of 2.9 percentage points
White students	Improvement: gain of 4.6 percentage points	Improvement -gain of 1.7 percentage points	No significant change in performance
Highest Quartile			
Black students	Improvement: gain of 3.0 percentage points	No significant change in performance	Improvement: gain of 2.6 percentage points
White students	No significant change in performance	Significant decline: 2.4 percentage points	Significant decline: 3.3 percentage points
<u>13-Year-Olds in the 8th Grade</u>			
Lowest Quartile			
Black students	Improvement: gain of 3.5 percentage points	No significant change in performance	Improvement: gain of 2.6 percentage points
White students	Improvement: gain of 1.5 percentage points	Improvement: gain of 2.0 percentage points	No significant change in performance
Highest Quartile			
Black students	Improvement: gain of 2.5 percentage points	No significant change in performance	Significant decline: 2.5 percentage points
White students	No significant change in performance	Significant decline: 4.1 percentage points	Significant decline: 3.2 percentage points
<u>17-Year-Olds in the 11th Grade</u>			
Lowest Quartile			
Black students	No significant change in performance	No significant change in performance	Improvement: gain of 1.6 percentage points
White students	Significant decline: 1.7 percentage points	No significant change in performance	Significant decline: 1.8 percentage points
Highest Quartile			
Black students	No significant change in performance	Significant decline: 3.9 percentage points	Significant decline: 5.5 percentage points
White students	No significant change in performance	Significant decline: 4.2 percentage points	Significant decline. 4.3 percentage points

SOURCE: National Assessment of Educational Progress, "Educational Winners and Losers, the Whos and Possible Why," (press release. February [illegible], 1983).

Among 13-year-olds as well, the lowest quartile showed mostly improvements in performance, albeit typically smaller than among the younger children. (This, too, is expected in light of the cohort pattern.) White students in the highest quartile again showed declines in two of three subjects; among blacks in this quartile, gains and losses were approximately balanced.

A similar discrepancy between the highest and lowest quartiles also appeared among the 17-year-olds, although overall--as expected--declines predominated over gains. Blacks in the lowest quartile showed no change in two subjects and a small gain in a third. Their white counterparts showed slight declines in two of three subjects. In contrast, in the top quartile, both races showed large declines in two subject areas.

Other Data

Data from other sources, however, are partially inconsistent with the NAEP data and call into question whether there was a general closing of the gap between high- and low-achieving students on a variety of tests and over the entire period of the achievement decline.

Tabulations of SAT candidates categorized by self-reported class rank show a similar narrowing of the gap between high- and low-achieving students since 1975. Moreover, this pattern occurred over most of the range of achievement; each group declined relative to all others ranking lower, bringing the scores of high-ranking and low-ranking students closer to each other. (These data unfortunately do not include reliable information about the bottom 20 percent.)

Ambiguous evidence on the relative trends among students in the top and bottom quartiles is found in the "Illinois Decade Study," a comparison of scores on a fairly high-level achievement test administered to Illinois high school juniors in the 1970 and 1981 school years. Declines in raw scores were consistently larger among students at the 75th percentile, albeit sometimes by a very small margin (see Table D-2).[2] On the other hand,

[2] *Student Achievement in Illinois, 1970 and 1981* (Springfield: Illinois State Board of Education, September 1983). Note that these data are not entirely comparable to the NAEP achievement subgroups analysis. Rather than reporting the average scores of all students above the 75th percentile--as in the NAEP reports--the Illinois Decade study reports results for the students at the 75th percentile. The same distinction applies to the data on scores at the 25th percentile. Thus, the NAEP analyses incorporate students who are further apart in their levels of achievement.

TABLE D-2. CHANGE ON THE ILLINOIS DECADE TEST AMONG STUDENTS AT THE 25th and 75th PERCENTILE

	75th Percentile		25th Percentile	
	Raw Change	Percent Change	Raw Change	Percent Change
Mathematics 1	-0.7	-4.6	-0.2	-2.7
Mathematics 2	-1.5	-12.5	-0.8	-13.1
English 1	-3.4	-16.0	-1.8	-13.8
English 2	-3.1	-15.2	-2.9	-22.1
Social Studies	-2.6	-15.0	-1.0	-11.0
Natural Science	-0.8	-6.6	-0.6	-3.8

SOURCE: CBO calculations based on Illinois State Board of Education, *Student Achievement in Illinois, 1970 and 1981*, Exhibit A-5; and J. Fyans, personal communication.

when the changes are expressed in proportional terms, this pattern disappears. The percent change in scores at the 25th percentile were sometimes smaller but sometimes larger than those at the 75th percentile.

Data from other tests, however, and from the SAT earlier in the period of decline (before 1975), cast doubt on the NAEP results. A tabulation of changes in SAT scores among groups of students divided by their percentile rankings on the SAT itself showed no comparable narrowing of the gap in the years before 1975. Indeed, in mathematics, the gap appears to have widened slightly (see the section below on "select students"). In addition, if the gap between the top and bottom quartiles were narrowing, one would expect a shrinking standard deviation--that is, a narrower distribution of scores. [3] Since the beginning of the 1970s,

[3] The standard deviation would shrink unless there were other, offsetting changes in the distribution of scores--such as a change in the distribution of scores in the middle two quartiles. Moreover, without such other distributional shifts, changes in the composition of the test-taking group would not alter this link between the standard deviation and the gap between the top and bottom quartiles. Any change in the standard deviation attributable to compositional changes (such as an increase resulting from lower dropout rates) would also be reflected in the gap between high- and low-achieving students.

however, both the SAT and ACT have shown stable or slightly increasing standard deviations. 4/ The standard deviation of scores on the ITBS has also been increasing. 5/ Between the 1970 and 1977 school years, the standard deviations of the SRA achievement series showed different changes, depending on subject area and grade. In general, they tended to increase in the younger grades but decrease in the higher grades. 6/ Given known problems in obtaining truly representative norming samples for such tests in different years, however, as well as changes in the representativeness of the samples over time, changes in the standard deviations of norming data should perhaps be given less weight than those in the other data sources. 7/

TRENDS AMONG COLLEGE-BOUND STUDENTS

Much of the public awareness of the achievement decline stems from the decline in SAT scores. But students taking college admissions tests (the SAT and ACT) and those planning to attend four-year colleges constitute only roughly half of the senior class, and their average level of achievement is above the overall average. 8/ Thus, it is important to gauge whether

4. The College Board, *College-Bound Seniors, 1984;* and American College Testing Program, unpublished tabulations.

5 H. D. Hoover, Iowa Testing Programs, personal communication, March 1984.

6. Science Research Associates, *SRA Achievement Series, Technical Report #3,* Table 2.

7. With respect to the problems in norming samples for such tests, see Roger F. Baglin, "Does 'Nationally' Normed Really Mean Nationally?" *Journal of Educational Measurement,* vol. 18 (Summer 1981), pp. 97-108; and Science Research Associates, *SRA Achievement Series, Technical Report #3.*

8. The group taking college-admissions tests and those entering college are not entirely the same, since not all college-bound students take the tests. In 1984, about 28 percent of those students graduating (excluding those obtaining high-school equivalency credentials) took the ACT, and 37 percent took the SAT. Those groups overlap to some unknown degree, however, so the proportion taking one or the other is less than the sum. The proportion taking such tests was lower during the early years of the decline. Similarly, 46 percent of all seniors in the class of 1980 (a larger group than all graduates, because of senior-year drop-outs) planned to attend at least four-year colleges. See The College Entrance Examination Board, *National College-Bound Seniors, 1985* (New York: The College Board, 1985); American College Testing Program, *Executive Summary: National ACT Assessment Results, 1984-1985* (Iowa City: ACT, 1985); National Center for Education Statistics, *Projections of Education Statistics to 1990-91* (Washington, D.C.: NCES, 1982); and Donald A. Rock, Ruth B. Eckstrom, Margaret E. Goertz, Thomas L. Hilton, and Judith Pollack, *Factors Associated with Decline of Test Scores of High School Seniors, 1972 to 1980* (Washington, D.C.: Center for Statistics, U.S. Department of Education, 1985).

trends on college-admissions tests are indicative of comparable trends among high-school seniors in general and, if not, whether differences reflect different trends among college-bound achievement subgroups or some other factors.

A difference between the trends shown by college admissions tests and tests given to all students need not indicate that achievement trends in the relatively high-achieving group of students taking the test are different from those in other achievement subgroups. A difference in score trends could also reflect changes in the self-selection of students taking the tests, or differences between the tests themselves and other tests administered to the student body as a whole.

As noted in Chapter IV, the decline in average scores on both the SAT and ACT were exacerbated by changes in the self-selection of students choosing to take the tests. In the case of the SAT, research suggests that over half of the decline between 1963 and 1970, but relatively little of it since then, reflected changes in the composition of the group taking the test.[9] Thus, in one sense, both the SAT and the ACT exaggerate the decline, in that the drop in average scores would have been substantially less if the test-taking group had remained constant or had changed only as the entire school-age population changed. (The research on this issue is described in CBO's forthcoming volume, *Educational Achievement: Explanations and Implications of Recent Trends.*)

This exaggeration of the decline, however, does not imply a greater drop in achievement among the relatively high-scoring achievement subgroups that tend to take these tests. A larger real decline in that group would be indicated if the decline on the SAT were larger than that on tests given to all high-school seniors, even after removing the influence of self-selection changes and accounting for differences between the tests. No existing studies, however, fully clarify whether there would be a greater decline on the SAT under those conditions, in part because there is not sufficient information to adjust for differences between the tests.[10]

[9] Advisory Panel on the Scholastic Aptitude Test Score Decline, *On Further Examination* (New York: The College Board, 1977), p. 18.

[10] In this context, one would want either confirmation that the tests involved would show similar trends if administered to the same students, or sufficient information to adjust the trends from one test to parallel those that would be produced by the other. Although equating studies that permit comparison of scores among tests at any one time are common, similar studies that permit comparisons of trends are largely lacking. Thus, as noted in Chapter III, much of the variation in trends among tests cited in this paper remains unexplained.

The available evidence, while not fully conclusive, does not suggest that the achievement decline was sharper among college-bound students than in the student population as a whole. Indeed, the decline might have been less severe in some college-bound groups during the early years of the decline. One study that directly compared trends in reading achievement among all seniors, college entrants, and SAT candidates between 1960 and 1972 found that the scores of college entrants, unlike those of SAT candidates, dropped only approximately as much as those of all seniors.[11] Since the college-bound population was also becoming less select during this period, the similarity might indicate that the average scores of some groups that traditionally sent many students to college were declining less than others, thereby offsetting the effects of the growing number of lower-achieving students going to college.[12]

For the years since 1972--the larger part of the period of decline on the SAT--there is no evidence that trends among college-bound students as a whole differed substantially in either direction from those among all seniors. In the nationally representative comparison of the NLS and HSB, seniors stating that they planned to attend four-year colleges or graduate schools showed declines in vocabulary, reading, and mathematics roughly comparable to those of the whole senior class.[13] Comparisons of trends on a variety of tests administered to juniors and seniors show some trends in the general student body that are more favorable than those on the SAT and ACT but others that are less favorable. Moreover, the trends on the SAT and ACT are inconsistent with each other (see, for example, Table III-2 in Chapter III). Given this inconsistency and the unexplained variation in trends among tests, disparities between the ACT and SAT and any given test administered to the student population as a whole could be reasonably attributed to differences in test characteristics rather than to variations in trends among achievement subgroups.

[11] Albert E. Beaton, Thomas L. Hilton, and William B. Schrader, *Changes in the Verbal Abilities of High School Seniors, College Entrants, and SAT Candidates Between 1960-1970* (New York: The College Board, 1977).

[12] Advisory Panel on the Scholastic Aptitude Test Score Decline, *On Further Examination*, pp. 13-16. Note that the SAT candidate group underwent changes in composition beyond those affecting the college-bound group as a whole, reflecting a change in the proportion and characteristics of those college-bound students taking the SAT.

[13] Donald Rock and others, *Factors Associated with Decline of Test Scores.*

SELECT STUDENTS

Because recent trends in the National Assessment have been relatively unfavorable among the top quartile of students, some people might assume that select students, variously defined, have also lost ground relative to other students.14/ This seems not to be the case, however. While some data show comparatively steep declines among select students, the available data as a whole do not, and the recent upturn appears to have been, if anything, particularly striking among some select groups. In addition, in mathematics--an area of particular public concern in recent years--select students might have been gaining ground for a considerable time.

Reports of trends among select students vary markedly, however. Some show greater declines than among other groups, while others show less marked declines or even no decline at all. This variation probably reflects the diversity both in criteria used to delineate select students and in the tests administered to them, as well as the sparseness of the available data. For example, the groups chosen to represent the select include: students scoring above specified thresholds on the SAT; students taking more selective tests, such as the College Board achievement and advanced placement tests; students in the highest ranks of their classes; and students taking certain advanced courses (such as high school calculus).

In addition, limitations of the data seriously cloud comparisons between select students and others. Only a few tests have been tabulated in a way that permits direct comparison of select and other students.15/ Those that are directly comparable are limited to high school students or, more narrowly, to college-bound juniors and seniors. Moreover, many of the tests that are designed intentionally for select students--such as the College Board achievement and advanced placement tests--are optional, and there is only limited information about the effects of changes in the test-taking groups on average scores. For example, the proportion of students taking

[14] Reports of trends among these students have used a variety of terms to label them. "Select students" is used here as a generic term for various groups of the highest-achieving students.

[15] Scores on many tests could be tabulated to permit such a comparison, subject to the limitations that small sample sizes and problems of scaling often would impose on how select a group could be assessed with confidence. Such reanalysis of the data at the level of individual students, however, is beyond the scope of this paper.

advanced placement tests has changed dramatically in recent years, as has their geographical distribution and the colleges they subsequently attend.

The SAT

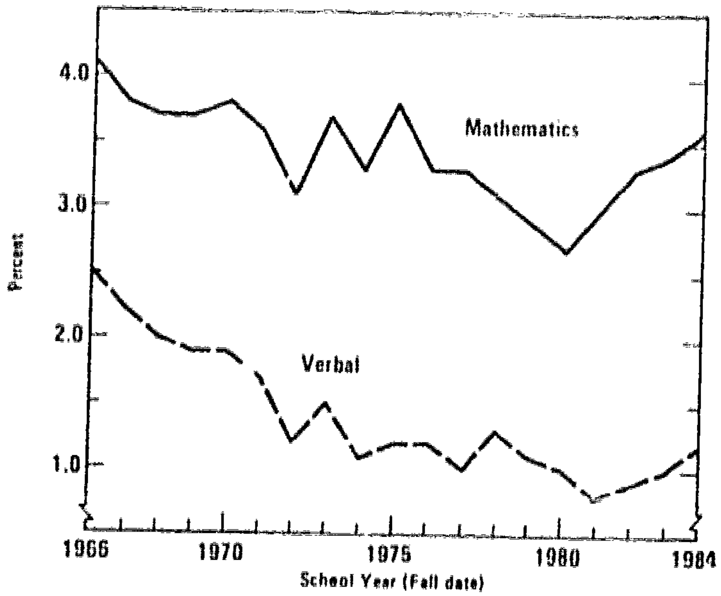
Perhaps the most commonly cited evidence of declining achievement among select students is the drop in the proportion of SAT candidates receiving very high scores. For example, the proportion receiving scores over 700 dropped sharply between 1966 and 1980, particularly on the verbal test (see Figure D-1). In 1966, roughly 2.5 percent of SAT candidates obtained verbal scores in excess of 700; that percentage had dropped to about 0.8 percent 15 years later. The drop was both more erratic and less severe on the mathematics test--from roughly 4.1 percent to 2.7 percent. (This parallels the fact that the drop in the mean score was much smaller on the mathematics test; see Chapter 3, Figure III-4.) A tabulation of this sort, however, cannot be compared directly with the overall decline, which is usually measured in terms of changes in the average scores themselves.

Trends in the proportion of candidates receiving high SAT scores also provide clear evidence that the recent upturn has been particularly sharp among some select students, at least in mathematics. The proportion of SAT-M scores over 700, for example, has risen roughly two-thirds of the way to its 1966 high level, even though it has been rising for only four years (see Figure D-1). The corresponding increase in the proportion of verbal scores over 700, however, has shown far less improvement.

Two other tabulations of SAT scores that are more directly comparable to common measures of the overall decline yield apparently--but perhaps not truly--contradictory information on the relative trends among select students. Both tabulations examine changes in the average scores of various select groups--rather than the number of students scoring above certain thresholds--but they use different criteria for categorizing students as select and encompass different time periods.

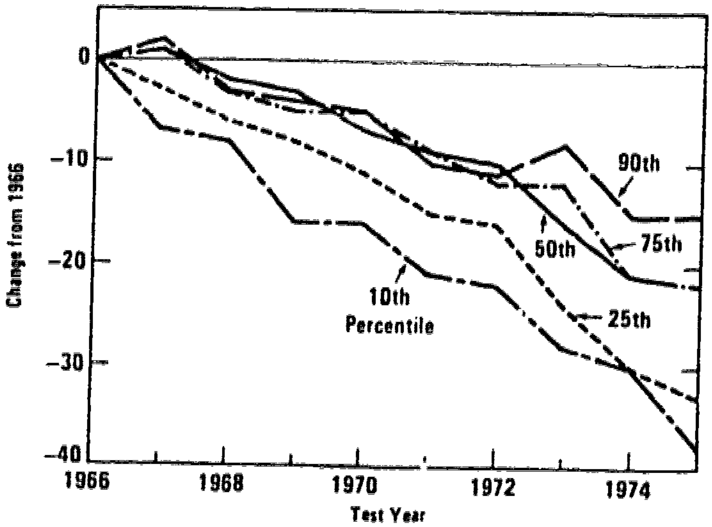
The first of these tabulations of select SAT scores indicates that from 1966 to 1975--a period that encompasses the worst of the SAT decline--average scores on the mathematics test declined somewhat less among the high-scoring than among lower-scoring SAT candidates (see Figure D-2). The average score at the 90th percentile declined the least, and scores at the 75th and 50th percentiles dropped substantially less than scores at lower percentiles. Only in the mid-1970s, however, did the top-scoring group show a different trend than that of the median SAT candidate. Moreover, no

Figure D-1.
Percent of SAT Scores Above 700 (By subject)



SOURCES: CBO calculations based on Hunter M. Breland, *The SAT Score Decline: A Summary of Related Research* (New York: The College Board, 1976); and the College Entrance Examination Board, *National College-Bound Seniors* (New York: The College Board, various years).

Figure D-2.
SAT Mathematics Scores for Selected Percentiles (Differences from 1966)



SOURCE: CBO calculations based on June Stern, *Selected Percentiles for Scholastic Aptitude Test Scores (1966-67 through 1975-76)* (New York: College Entrance Examination Board, 1977).

similar differentiation appeared on the verbal scale. 16/ Unfortunately, no comparable tabulation of SAT scores is available for years after 1975.

The second tabulation, which only began in 1975 and in which select students were included on the basis of self-reported class rank rather than SAT scores, shows virtually the opposite pattern: more select groups lost ground on the SAT verbal test relative to other students (see Figure D-3). 17/ This trend was apparent both during the last years of the decline and during the first few years of the subsequent upturn. In contrast, since 1982, the gap between the various groups has largely remained constant. Indeed, this pattern was not limited to select students; across the entire range, students with higher class rank showed less favorable trends than did students with lower class rank. 18/ Scores of students reporting themselves to be above the 90th percentile in class rank fell 16 points on the SAT-V between the 1975 and 1982 school years and only began turning up in 1983. The pattern among students between the 80th and 90th percentiles is quite similar, but the decline is four points smaller, and the subsequent upturn is clearer and might have begun a few years earlier. In contrast, the average scores of the broad middle of students--those falling between the 20th and 80th percentiles in class rank--showed at most a small drop between 1975 and 1979 and have been rising quite steadily since.

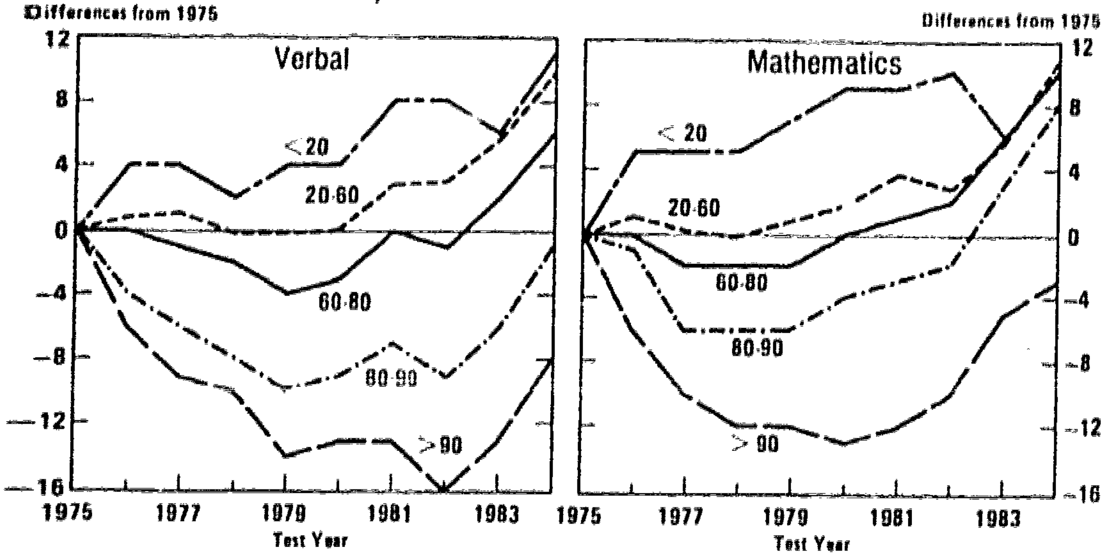
While less favorable trends appeared among students with higher class ranks on the SAT mathematics scale as well, the mathematics trends differed in some respects (see Figure D-3). As in the case of the verbal scale, the widening gap between achievement groups was quite consistent across the entire range of achievement levels, and the upturn began consistently earlier in lower-ranked groups of students. In the case of mathematics, however, the widening of the gap between high- and low-achieving groups had ended before the overall rise in scores began in 1981, and, indeed, the top 10 percent of students gained a bit relative to others during the first years of the score increase.

16. June Stern, *Selected Percentiles for Scholastic Aptitude Test Scores (1966-67 through 1975-76)* (New York: The College Board, 1977).

17. William W. Turnbull, *Changes in SAT Scores: What Can They Teach Us?* (College Board - ETS Joint Staff Research and Development Committee, forthcoming), Table II.

18. Although the trend among students below the 20th percentile is largely consistent with this generalization, it cannot be interpreted with confidence, for it reflects very few students--only 0.6 percent of SAT candidates in the 1983-1984 school year.

Figure D-3.
SAT Scores by Percentile of Class Rank (By subject, differences from 1975)



SOURCE: CBO calculations based on the College Entrance Examination Board, *National College-Bound Seniors* (New York: The College Board, various years).

Is the apparently steeper SAT decline after 1975 among students with high class ranks inconsistent with the comparable or even lesser declines of students with high SAT scores in earlier years? Not necessarily. The mathematics trends suggest that the upturn might have begun earlier among lower-achieving students. If so, it could cause an apparently greater decline among select students during the last few years of the decline even if select students showed comparable or lesser drops over the entire period of the decline. In addition, select students might have declined less on the SAT during the earlier and middle years of the decline but more at its end--a pattern that could easily arise if the trends reflect a variety of different causes. Alternatively, class rank and SAT percentiles might delineate different select groups that experienced different trends throughout the decline. This possibility is strengthened by the fact that class rank, unlike SAT percentiles, is based on self-reports by students and is therefore subject not only to random error, but also to systematic differences in response bias among different groups of students. Finally, changes in grading criteria or students' choices of classes might have altered the meaning of class rank. Those students currently ranking in the top 10 percent, for example, might be dissimilar in some respects from those with comparable ranks in 1975.

The SAT trends among students divided by class rank also fail to show the sharp relative gains in mathematics scores among select students evidenced by the dramatic rise in SAT-M scores above 700, but this discrepancy might also be more apparent than real. The difference suggests that the atypically sharp rise in mathematics achievement is limited to a more select group of students than those reporting themselves in the top 10 percent of their classes. Students scoring over 700 are far fewer in number than those reporting themselves to be in the top 10 percentile of class ranks; even after the recent increases, only 3.6 percent of SAT candidates are in the former group, compared with 21.1 percent in the latter. 19/ The former group also presumably comprises students who are more select in terms of their coursework in mathematics.

The Illinois Decade Study

This study suggests that the decline among select students was no worse, and perhaps slightly less severe, than that among other students. The decline among students at the 95th percentile (that is, those at the cutoff for the top 5 percent) was generally similar to that of students at the 75th and 50th percentiles, with one exception: on one of two mathematics tests, those students at the 95th percentile showed almost no decline. 20/

The Iowa Test of Basic Skills

National norming data from the ITBS show scores of eighth-grade students at the 90th percentile declining considerably less than scores of the median student between 1970 and 1977--a period that includes the first year of the upturn. On this test, unlike some of the others described here, the relative gains of the select students were greater in language-related areas than in mathematics. 21/

19. The College Board, *National College-Bound Seniors (1985)*, Tables 1 and 7.

20. On that particular math test, the lowest-scoring students (in this case, those at the 25th percentile) declined by as little as those at the 95th percentile in absolute terms, while those students falling in between declined substantially more. Illinois State Board of Education, *Student Achievement in Illinois*, p. 10.

21. Hieronymus, Linquist, and Hoover, *Iowa Test of Basic Skills: Manual for School Administrators*, Table 6.24. Similar patterns were apparent at many of the other grade levels as well, but their interpretation is less clear, since the differences at younger ages included more of the period of increasing scores.

The Second International Assessment of Mathematics

A recent international assessment of mathematics achievement suggests that select American students--in this case, those taking calculus while in high school--have improved in mathematics. This assessment, carried out in 1981-1982 in a national sample of American schools, included testing of seniors in calculus and pre-calculus classes--together, about 10 percent to 12 percent of seniors. The performance of this group was slightly superior to that of comparable students in a similar international assessment 17 years earlier (based on items included in both assessments), although it was still quite poor by international standards. This improvement appears to have been far stronger among the students in the calculus classes. 22/

The College Board Achievement Tests

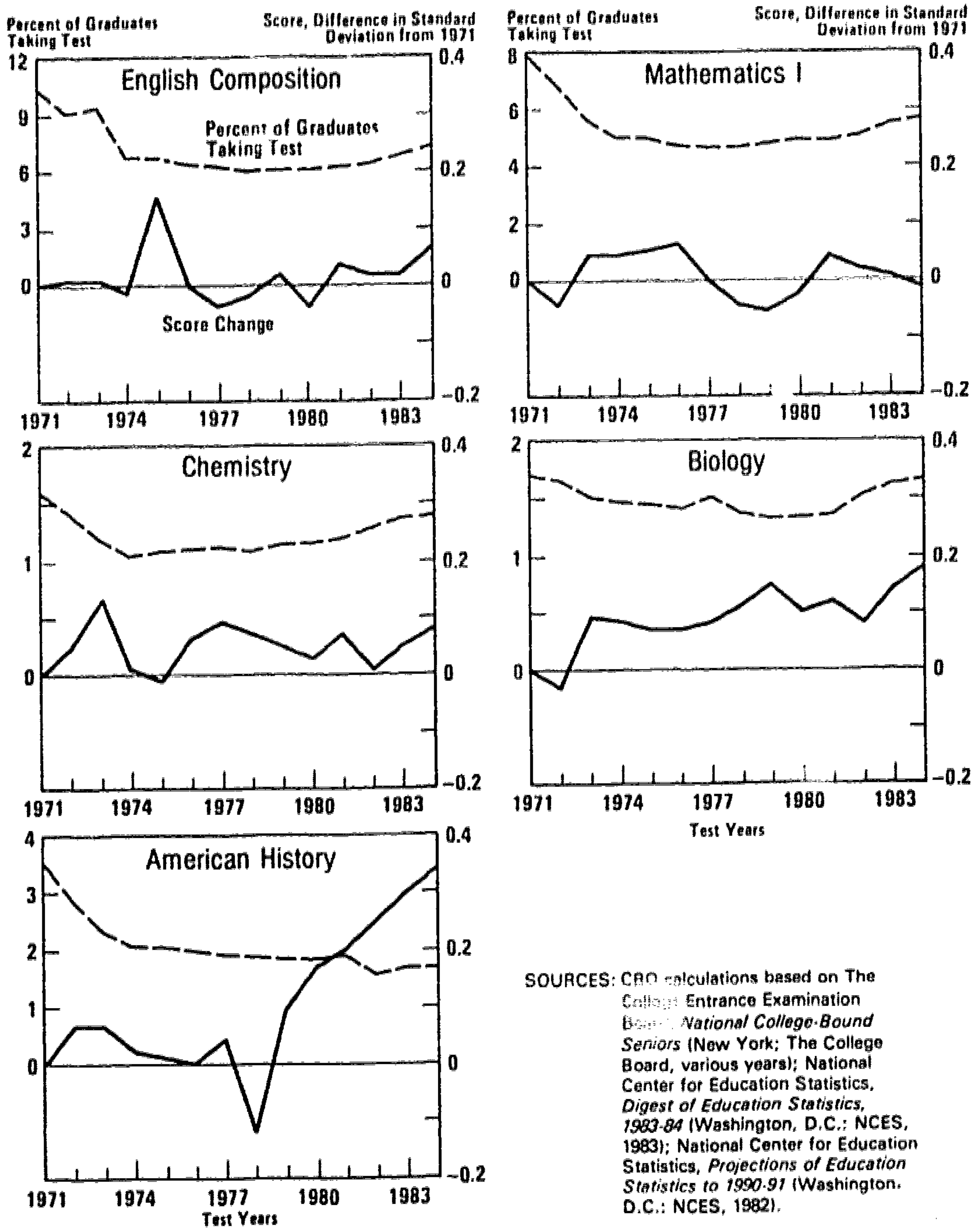
These tests of achievement in specific subject areas are taken by a small fraction of 1 percent to about 10 percent of graduates, depending on the subject area and year. Typically, they showed stability or slight increases during the last half of the period of declining achievement, but this might merely reflect a rapid drop in the proportion of graduates taking the tests (see Figure D-4). 23/ That is, if the declining proportion of graduates taking

[22] F. Joe Crosswhite, John A. Dossey, Jane O. Swafford, Curtis C. McKnight, Thomas J. Cooney, and Kenneth J. Travers, *Second International Mathematics Study: Summary Report for the United States* (Champaign, Illinois: Stipes Publishing Co., 1985), pp. 63, 70-73. Details of the earlier assessment can be found in Torsten Husen, ed., *International Study of Achievement in Mathematics: A Comparison of Twelve Countries* (Stockholm and New York: Almqvist & Wiksell and John Wiley & Sons, 1967).

[23] Because of scaling, the drop in the proportion of students taking the achievement tests is more marked than it might seem in Figure D-4. For example, the proportion of students taking the biology test dropped by about 22 percent between 1971 and 1979, but that decline appears moderate in Figure D-4.

Test score data are from College Board, *National College-Bound Seniors*, various years. Comparable data on scores and participation rates are unavailable before the 1971 school year. Participation rates are obtained by the dividing the number of test takers in a given year by the number of high-school graduates in that year in *Projections of Education Statistics to 1990-91*, (Washington, D.C.: National Center for Education Statistics, 1982). This produces a slight overestimate of the proportion of graduates taking the test, because some students take the test in their junior year and repeat it the following year. More important, it overstates the selectivity of the tests in areas in which the SAT and the Achievement Tests are heavily used. It adds to the denominator students in areas where few students take these tests (for example, areas in which the ACT is the dominant college-entrance test).

Figure D-4.
ETS/CEEB Scores and Percent of Graduates Taking Test



SOURCES: CBO calculations based on The College Entrance Examination Board, *National College-Bound Seniors* (New York; The College Board, various years); National Center for Education Statistics, *Digest of Education Statistics, 1983-84* (Washington, D.C.: NCES, 1983); National Center for Education Statistics, *Projections of Education Statistics to 1990-91* (Washington, D.C.: NCES, 1982).

the tests reflects a drop in the number of less able students taking the tests, the resulting increase in the ability level of the remaining group taking the tests might have masked a decline within ability groups. For example, the proportion of graduates taking the English composition test dropped roughly from 10 to 6 percent between the 1971 and 1978 school years, and similar declines in participation occurred in other subject areas as well.

Conversely, the relative stability of many of the College Board achievement test scores since 1979 might hide a substantial increase in achievement within ability groups. Since 1979, average scores on the more common College Board achievement tests have generally held stable or increased modestly in the face of moderate-to-large increases in the proportion of students tested. (American History is an exception; it showed a large increase in average scores but a slight decrease in participation.)

The College Board's Advanced Placement Tests

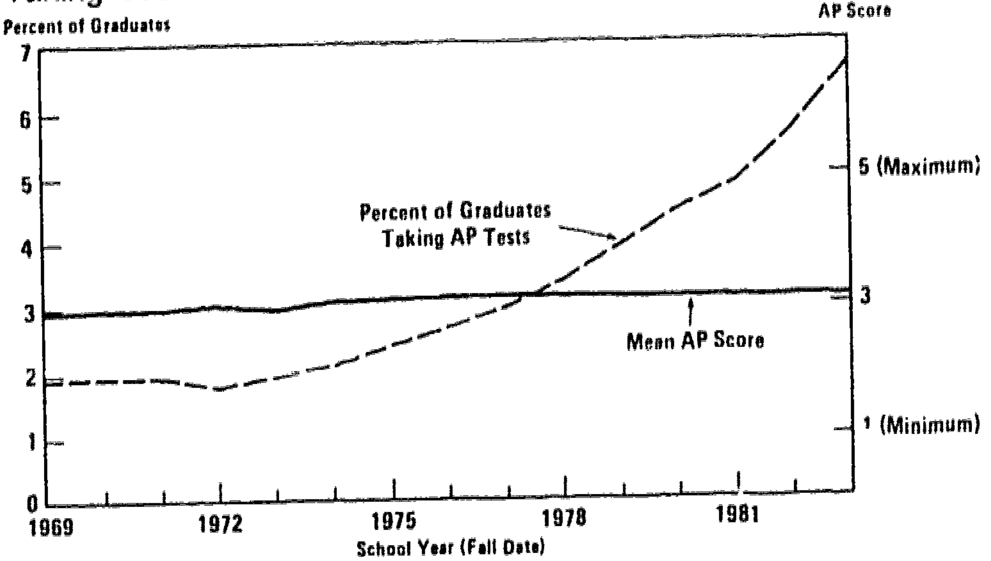
Average scores on this set of tests--taken by college-bound students seeking college credit for advanced coursework in high school--has remained stable since 1969. This stability, however, might mask a sizable increase in educational accomplishment.

Relatively few graduates take each of the Advanced Placement (AP) tests, but the total proportion taking any of them has roughly tripled--from under 2 percent to about 6 percent--over the past decade.[24] During this decade of rapid growth--as well as the preceding half-decade of fairly stable test volume--the average score on AP tests in all subjects remained quite stable, increasing about 5 percent (see Figure D-5).

The rapid growth in the proportion of seniors taking the AP tests need not indicate the sort of compositional changes that affected the SAT in the 1960s, and the stability of AP scores accordingly should be interpreted differently. In the case of the SAT, the growth in the proportion of students taking the test in part indicated an increase in the proportion of test takers from lower-ability groups. In such a situation, a stable overall average score would indicate increasing achievement within ability groups. In the

[24] Data are from published and unpublished College Board tabulations. These proportions are subject to the same caveats as are described above with respect to the College Board achievement tests.

Figure D-5.
Average Advanced Placement Scores and Percent of Graduates Taking Tests



SOURCES: CBO calculations based on Advanced Placement Program of the College Board, *Advanced Placement Yearbook, 1984* (New York: The College Board, 1984), and unpublished tabulations; National Center for Education Statistics, *Digest of Education Statistics, 1983-84* (Washington, D.C.: NCES, 1983); National Center for Education Statistics, *Projections of Education Statistics to 1990-91* (Washington, D.C.: NCES, 1982).

case of the AP tests, however, much of the growth in volume reflects the expansion of the AP program into additional geographic areas, as additional universities decided to offer credit for AP tests and more school districts and individual schools decided to offer advance courses prepariレg students for the AP tests. For example, the decision of some large state universities to offer AP courses contributed substantially to the growth of the AP program, and students going to such universities--such as the University of California, the University of North Carolina, and the State University of New York--now account for a large share of the total number of AP examinations. 25/ Thus, the growing proportion of students taking AP exams might be lowering the average ability of the test-taking group, but

25. College Entrance Examination Board, unpublished tabulations; and Harlan Hanson, The College Board, personal communication, March 1985.

probably far less than did the growth of the SAT pool two decades ago. While some of the new students added to the AP pool might be lower in ability than those in the smaller pool a decade ago--when more selective schools contributed a greater share of the students--many are probably comparable in ability, differing only in geographic location or family income.

The constancy of AP scores in the face of rapid growth in the number of test-takers accordingly can be seen as an increase in educational accomplishment. To the extent that the average ability of the pool might have decreased, the stable scores reflect an increase in the scores obtained by students at any given ability level. To the extent that additional students of comparable ability have been drawn into the program, the program's growth represents a dramatic increase in the advanced-level coursework of highly able students--that is, it can be seen as a growth in their educational attainment.

CONCLUSION

Taken together, the available data provide only spotty and inconsistent suggestions that achievement trends have been relatively more favorable in some achievement subgroups than in others. There are some indications of relative gains at both ends of the achievement scale--that is, among students in the lowest quartile and among certain select students. These signs, however, appear limited to certain tests. In addition, if these relative gains are not an artifact of certain aspects of those particular tests, some apparently might be confined to relatively short periods.

Indeed, the data suggest that generalizations about relative gains in various achievement subgroups are risky, and that inferences for educational policy might not be warranted. The variation in trends from one data source to another--and even from one tabulation to another of a single data source--appears more striking than any generalizations about relative trends among achievement subgroups. The uncertainty engendered by this variability is exacerbated by the many gaps in the available data and by technical problems entailed in using the data in their current form to draw conclusions about achievement subgroups.

APPENDIX E

DIFFERENCES IN ACHIEVEMENT TRENDS AMONG BLACK, HISPANIC, AND NONMINORITY STUDENTS

Evidence that the average scores of black and Hispanic students have risen relative to those of nonminority students--but remain well below them--is summarized in Chapter IV. Because that conclusion has considerable importance, the evidence underlying it is presented in more detail in this appendix.[1]

BLACK STUDENTS

Although data on differences in achievement between black and nonminority students at any one time are abundant, data sources showing relative trends in achievement in those two groups are surprisingly rare. In the course of this study, nine data sources with separate trend data for black and nonminority students were located. Two are nationally representative: the National Assessment of Educational Progress (NAEP), and a comparison of the National Longitudinal Study of the High School Senior Class of 1972 (NLS) and the High School and Beyond study (HSB).[2] Two others are national but unrepresentative: the Scholastic Aptitude Test (SAT) and American College Testing Program (ACT) tests. Data are also available from two statewide assessments (North Carolina and Texas) and three local districts (Houston, Texas; Cleveland, Ohio; and Montgomery County, Maryland).

[1] For an explanation of the ethnic classifications used in this paper, see Chapter IV. The classifications used in the data sources cited here are not entirely consistent. In each case, the scores of black students have been compared with the group which comes closest to being "nonminority"--that is, the group that excludes the largest share of identified minority groups. This nonminority group, however, varies among data sources. The SAT "white" category, for example, specifically excludes Asian Americans, native Americans, Puerto Ricans, and Mexican Americans. In contrast, the closest comparable category in the NLS/HSB comparison combines non-Hispanic whites with Asian-American and native American students.

[2] Donald A. Rock, Ruth B. Ekstrom, Margaret E. Goertz, Thomas L. Hilton, and Judith Pollack, *Factors Associated with Decline of Test Scores of High School Seniors, 1972-1980* (Washington, D.C.: Center for Statistics, U.S. Department of Education, 1985).

Eight of these nine data sources showed a consistent and unambiguous narrowing of the gap between black and nonminority students, leaving little doubt that this pattern is real and not an artifact of some aspects of the tests or groups tested. The one partial exception is the ACT. That test did show a small narrowing of the gap, but the evidence is somewhat questionable because of inconsistencies among subject areas and large year-to-year fluctuations. While the reasons for that one partial anomaly are not clear (several possible explanations are discussed below), it is not sufficient to call the convergence of scores on all of the other eight tests into serious doubt. The consistency among the other eight tests is particularly persuasive in the light of the variation in grade levels, test characteristics, and student characteristics from one test to another.

This convergence in the scores of black and nonminority students appears to have three components. The scores of black students:

- Declined less that those of nonminority students during the later years of the general decline;
- Stopped declining, or began increasing again, earlier; and
- Rose at a faster rate after the general upturn in achievement began.

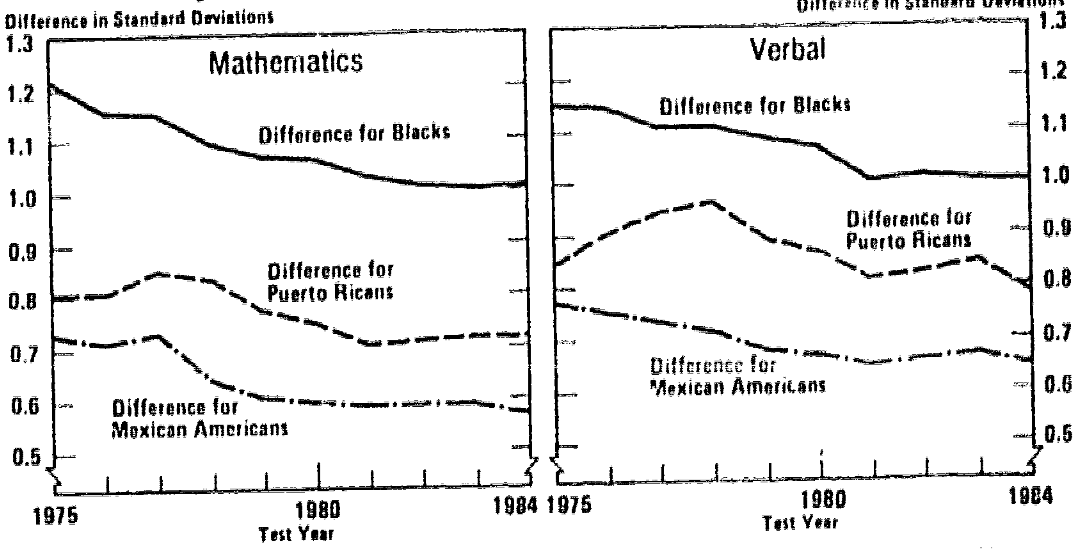
These specific conclusions, however, are less certain than is the overall convergence between the two groups, for not all are apparent in all eight of the data sources.

The SAT

Since 1975, black students have gained relative to nonminority students on both scales of the SAT (see Figure E-1)--a trend that ended with the 1981 and 1983 school years (on the verbal and mathematics scales, respectively). During the late 1970s, while nonminority students continued to lose ground, black students improved their scores on the mathematics scale and held about constant on the verbai scale. During the first years of the overall upturn in scores, blacks gained more rapidly than nonminority students.

Both the size of the gap and the rate at which it has been shrinking can be gauged by comparing the average SAT scores of black students with the distribution of scores of nonminority students. In 1975, the average black student's score corresponded roughly to the 11th and 12th percentiles among nonminority students on the mathematics and verbal scales,

Figure E-1.
Minority/Nonminority Differences on the SAT (In standard deviations, by subject)



SOURCES: CBO calculations based on "College Board Data Show Class of '85 Doing Better on SAT, Other Measures of Educational Attainment" (press release, The College Board, 1985), and Solomon Arbeiter, *Profiles, College-Bound Seniors, 1984* (New York: The College Board, 1984).

NOTE: Plotted points are the differences in standard deviations between the mean score of each group and the mean score of nonminority students.

respectively. In 1984, the average black scores had risen to about the 16th percentile among nonminority scores on both scales. 3/ While this change might appear slight, the annual rate of change is in fact roughly comparable to the average rate of the total SAT decline--a trend that few would label insignificant. 4/

3. These estimates are based on nonminority (white) within-group standard deviations in 1983-1984 reported in Solomon Arbeiter, *Profiles, College-Bound Seniors, 1984* (New York: The College Board, 1984), p. 81. Although the within-group standard deviation is technically the appropriate index in a comparison of this sort, using the more commonly available total standard deviation does not substantially alter the results. Moreover, the standard deviations of most tests have changed only very slowly, so the choice of a year from which to take a standard deviation is largely immaterial.

4. During the total period of decline, average SAT verbal and mathematics scores declined at annual rates of 0.028 and 0.016 standard deviations per year, respectively. During the past nine years (the only period for which data are available), the gap between black and nonminority students has shrunk at annual rates of 0.017 and 0.023 standard deviations per year on the verbal and mathematics scales, respectively (based on 1983-1984 standard deviations in the total SAT sample).

The ACT

Black students have gained relative to others on the ACT composite scale since 1970, but that gain has been small and is overshadowed by large year-to-year fluctuations in the size of the gap (see Figure E-2). In addition, the trend has been inconsistent from one subject area to another. The gaps appear to have narrowed in Social Studies and English, for example, while widening in mathematics. 5/

These anomalous patterns on the ACT have a number of possible explanations. For example, the year-to-year instability of the trends might reflect fluctuations in the sample (10 percent of all those taking the test). The relatively small change in the gap over the total time period might in part reflect the nonrepresentative group of students taking the ACT. It might also reflect the fact that in this case the comparison is with all non-black students--a group that includes, for example, Hispanic students and a small number of Asian and native American students. Since Hispanics appear also to have been gaining on nonminority students, trends in this nonblack group might have been slightly more favorable than among non-minority students, leading to a slight understatement of the relative gains of black students.

The NLS and HSB

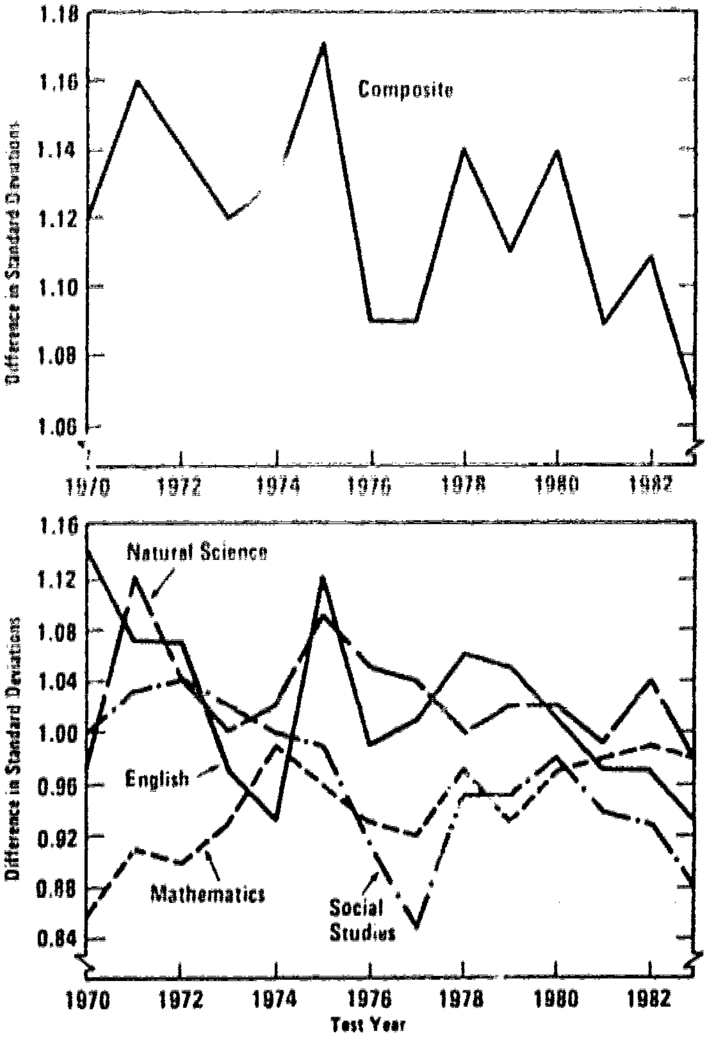
The narrowing of the gap between black and nonminority students is apparent also in the nationally representative comparison of the graduating classes of 1972 and 1980 (school years 1971 and 1979) based on the NLS and HSB studies. 6/ In this instance, however, as in the case of the ACT, the trends are clouded by the inclusion of several minority groups in the same category as nonminority students. In all three subjects tested--reading,

Many ACT tabulations provide the scores of black students and all students, but not those of nonblack students. If the scores of all students are used as a comparison instead of those of nonblack students, the gap appears--spuriously--to have been shrinking a bit faster than these figures suggest. The reason for the discrepancy is that an increase the proportion of black students (in the particular sample used) during the late 1970s d early 1980s lowered the scores of the total group relative to the nonblack group.

The ACT data in Figures E-2 were calculated using the 1977 total-sample standard deviations. Using more recent standard deviations does not alter the results appreciably, and substituting within-group nonblack standard deviations should have only a small effect.

Donald A. Rock and others, *Factors Associated with Test Score Decline.*

Figure E-2.
Black/Nonblack Differences on the ACT (In standard deviations, by subject)



SOURCES: CBO calculations based on: American College Testing Program, unpublished and undated tabulations; American College Testing Program, "Overview of Selected Results" (ACT, unpublished and undated material); Jackie Woods, ACT, personal communication, December 1985.

vocabulary, and mathematics--the largest average declines occurred among a group comprising non-Hispanic whites, Asians, and American Indians (but dominated by the far more numerous non-Hispanic whites.) Trends among black students ranged from a small gain in mathematics to a larger but modest decline in reading (see Table E-1). 7/

7. None of these changes in the average scores of black students was statistically significantly different from no change. See Rock and others, *Factors Associated with Test Score Decline*, Tables D-1, D-2, and D-3.

TABLE E-1. AVERAGE ACHIEVEMENT OF BLACK AND OTHER STUDENTS IN THE NLS AND HSB, BY SUBJECT

Category	1972	1980	Change
Vocabulary			
Black	3.28	3.20	-0.08
Other [a]	7.04	6.22	-0.82[b]
Reading			
Black	5.94	5.56	-0.38
Other [a]	10.51	9.57	-0.94[b]
Mathematics			
Black	6.50	6.69	0.19
Other [a]	13.90	12.97	-0.93[b]

SOURCE: Rock and others, *Factors Associated with Decline of Test Scores*, Tables D-1, D-2, and D-3.

[a] "Other" category includes non-Hispanic whites, Asian Americans, and American Indians.

[b] Statistically significant at the .05 level or less.

The National Assessment of Educational Progress (NAEP)

The gap between black and nonminority students also narrowed at all three ages tested in the NAEP (see Tables E-2 and E-3). Moreover, this narrowing appeared quite consistently in both the top and bottom achievement quartiles (see Table D-1 in Appendix D). In some cases, both groups lost ground, but nonminority students lost more; in others, both blacks and nonminority students gained, but blacks gained more. In some instances, black scores increased while the nonminority average declined. Although not presented in detail here, NAEP assessments in the areas of social studies and writing also showed a narrowing of the gap among 9- and 13-year-olds.

TABLE E-2. READING PERFORMANCE OF BLACK AND NONMINORITY STUDENTS IN THE NATIONAL ASSESSMENT (Average percent of items answered correctly and proficiency scores)

	1970	1974	1979	1983	Change 1970-1979
			Percent Correct		
Age 9					
Nonminority a/	66.4	67.0	69.3	NA	2.8
Black	49.7	54.5	59.6	NA	9.9
Age 13					
Nonminority a/	62.6	61.9	62.6	NA	.0
Black	45.4	46.5	49.6	NA	4.2
Age 17					
Nonminority a/	71.2	71.2	70.6	NA	-0.7
Black	51.7	52.1	52.2	NA	0.5
			Proficiency Scores		
Age 9					
Nonminority b/	214.4	215.9	219.7	220.1	5.7
Black	169.3	181.9	188.9	188.4	19.1
Age 13					
Nonminority b/	260.1	260.9	263.1	263.4	3.3
Black	220.3	224.4	231.9	236.8	16.5
Age 17					
Nonminority b/	290.4	290.7	291.0	294.6	4.2
Black	240.6	244.0	246.1	263.5	22.9

SOURCES: National Assessment of Educational Progress, *Three National Assessments of Reading: Changes in Performance, 1970-1980* (Denver: NAEP/Education Commission of the States, 1981), Tables A-1, A-5, and A-9, and *The Reading Report Card: Progress Toward Excellence in Our Schools* (Princeton: NAEP/Educational Testing Service, 1985), Data Appendix.

NOTE: NA denotes not available.

a. Includes Hispanics in all years. See footnote 9.
b. Includes Hispanics in 1970 only. See footnote 10.

TABLE E-3. MATHEMATICS PERFORMANCE OF BLACK AND NONMINORITY STUDENTS IN THE NATIONAL ASSESSMENT a/
(Average percentage of items answered correctly)

Group	1972 (Estimated) b/	1977	1981	Change 1972-1981
Age 9				
Nonminority	60.1	58.1	58.8	-1.28
Black	40.2	43.1	45.2	4.99
Age 13				
Nonminority	62.3	59.9	63.1	0.84
Black	41.1	41.7	48.2	7.07
Age 17				
Nonminority	66.7	63.2	63.1	-3.56
Black	46.3	43.7	45.0	-1.32

SOURCES: 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), Table 5.1; and CBO calculations based on National Assessment of Educational Progress, *Mathematical Technical Report: Summary Volume* (Denver: NAEP/Educational Commission of the States, 1980), Tables 2, 3, and 4.

a. Nonminority category excludes Hispanics in all years.

b. These estimates for 1972 differ from published NAEP results for the 1972 assessment. The published results for that year are based either on the 1972 item pool or on the items used in both 1972 and 1977, while the trend results comparing the 1977 and 1981 assessments reflect items used in both the 1977 and 1981 assessments. In order to circumvent the large disparities in the item sets, 1972 results were estimated here by adjusting the 1977 results (on the items used in 1977 and 1981) by the 1972-to-1977 change (on the items used in 1972 and 1977).

On the other hand, in science, no clear narrowing of the gap was apparent. 8/

The NAEP provides a somewhat different view than the SAT of the magnitude of the achievement gap between black and nonminority students and of the rate at which that difference is shrinking. The NAEP, in contrast to the SAT, is designed to assess the degree to which students have mastered commonly taught material. Moreover, until recently, the NAEP was scaled in a way that is intuitively clearer--albeit less useful in some important respects--than the SAT; scores are typically presented as the average percent of items answered correctly by a given group of students. In the early 1970s, black students on average correctly answered about a third fewer items in math and a fourth fewer in reading than did their nonminority peers. 9/ For example, nonminority nine-year-olds averaged 60 items correct in mathematics, compared with about 40 items answered correctly by the average black student. In proportional terms, these differences were quite similar in all three age groups tested.

Throughout the 1970s, differences between black and nonminority students in NAEP scores shrank more rapidly among elementary and junior-high students than among high school students. Among nine-year-olds, the average black student's mathematics score was roughly a fourth below the average nonminority score in 1981, compared with a third below in 1972. In reading, the average black score went from a fourth below the

8. See Nancy W. Burton and Lyle V. Jones, "Recent Trends in Achievement Levels of Black and White Youth," *Educational Researcher*, vol. 11 (April 1982), pp. 10-14, 17. Burton and Jones suggest that the racial gap has narrowed in science as well, but that change appears largely to be an artifact of differences in the content of the tests given in different pairs of years. When the 1972-1976 change in racial differences on the item set administered in both of those years is added to the 1969-1972 change on the set used in both of those years, the trend in the racial difference over the entire period considered is nearly zero. This can be seen from their Figures 4 and 5 and, more precisely, from Tables A-2, A-3, and A-4 in National Assessment of Educational Progress, *Three National Assessments of Science*.

9. In these reading data, Hispanics are included in the nonminority category (National Assessment of Educational Progress, *Three National Assessments of Reading*, p. 2). While including Hispanics in the nonminority category lowers the average score of that group, its effect on the trends is unclear. On the one hand, the relative gains of Hispanic students during that period--described subsequently--would make the trends in the nonminority group more favorable and thus attenuate the comparative gains among blacks. On the other hand, the growth of the Hispanic share of the school-age population would make trends in the nonminority group less favorable and thus exaggerate the relative gains of blacks. In contrast, in the mathematics data, Hispanic students are separated (National Assessment of Educational Progress, *Changes in Mathematical Achievement, 1973-78*, p. 29).

nonminority score in 1970 to less than 15 percent below in 1979. The gap narrowed slightly less among 13-year-olds and very little among 17-year-olds.

In the most recent (1983) reading assessment, NAEP scores are reported in terms of "proficiency scores" that permit comparison of the performance of students in different age groups--providing yet another way of gauging the gap between black and non-minority students. Through the 1979 assessment, the;e data reveal the same pattern noted above, with one addition--through 1979, black 17-year-olds were on average less proficient in reading than nonminority 13-year-olds (see Figure IV-5 in Chapter IV). 10/

Since 1979, these new NAEP data indicate that the closing of the gap between black and nonminority students accelerated among 17-year-olds while ending among nine-year-olds. (Because of the large gains among black 17-year-olds, the average performance in the groups reached the level of the average among nonminority 13-year-olds for the first time.) This pattern makes sense in terms of a cohort model; in both age groups, the black students born in the mid-1960s contributed the most marked gains (see Figure IV-5 in Chapter IV). On the other hand, these trends among 17-year-olds are inconsistent with the SAT data, which show the relative gains of black students ending in the last few years.

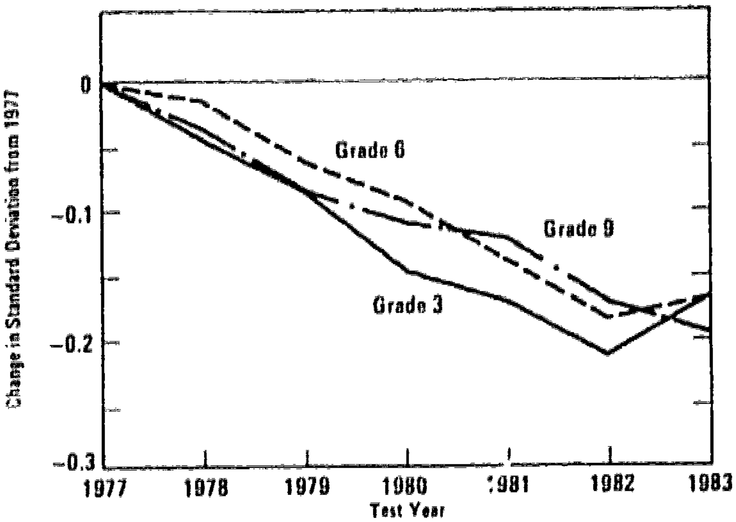
State-Level Data

Statewide assessments from two states, North Carolina and Texas, provide trend data separately for black and nonminority students, and both show a narrowing of the gap between the two groups. The North Carolina statewide assessment program provides average scores of black and white students on a standardized achievement test (the CAT) since 1977. In all three grades tested (3, 6, and 9), the gap has narrowed considerably (see Figure E-3).11/

[10] In these tabulations, Hispanics are included in the white (or nonminority) category only in 1970 (National Assessment of Educational Progress, *The Reading Report Card*, Data Appendix). Their being included only in the base year and excluded thereafter exaggerated the improvement among whites, thus attenuating the relative gains of black students.

[11] The trends in Figure E-3 were calculated using the total standard deviation from the 1977 norming sample for the California Achievement Tests (*California Achievement Tests, Forms C and D, Technical Bulletin 1* (Monterey: CTB/McGraw-Hill, 1978), Table 8). If standard deviations based on the North Carolina data were available, their use would have altered the specific numbers in Figure E-3, but the differences most likely would have been relatively small, and the convergence of black and white students' scores would still be apparent.

Figure E-3.
North Carolina Black/White Differences on CAT (Change from 1977 in standard deviations, by grade)



SOURCES: CBO calculations based on North Carolina State Department of Education, unpublished tabulations, and *California Achievement Tests, Form C and D: Technical Bulletin 1* (Monterey: CTB/McGraw-Hill, 1979).

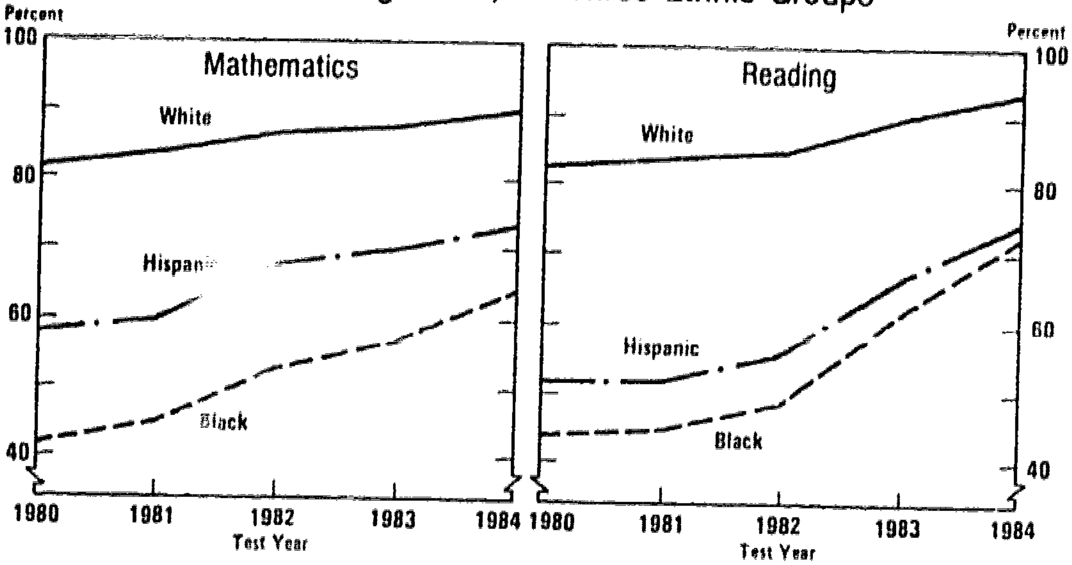
Black ninth-grade students have also improved their average achievement on the Texas statewide mathematics and reading tests more rapidly than have nonminority students during the few years for which data are available (see Figure E-4). 12/

HISPANIC STUDENTS

As noted in Chapter IV, trend data about Hispanic students are sparser than those about black students, and their meaning is clouded by inconsistencies

12. The Texas scores are tabulated as percentages of students in each group exceeding a specific criterion score. Since the proportion of white students exceeding the criterion is very high, the convergence of the scores of black and nonminority scores may in part reflect a "ceiling effect"--that is, the fact that the success rate among nonminority students cannot rise much more. Even after a mathematical correction of this problem (normalizing the proportions with a logit transformation), however, the gap appears to be narrowing appreciably, albeit at a slower rate than in the unadjusted data presented in Figure E-4.

Figure E-4.
Percentages of Grade-Nine Texas Students Passing Mathematics and Reading Tests, for Three Ethnic Groups



SOURCE: W. James Popham, Keith L. Cruse, Stuart C. Rankin, Paul D Sandifer, and Paul L. Williams, "Measurement-Driven Instruction: It's on the Road," *Phi Delta Kappan*, vol. 66 (May 1985), pp. 628-634.

in the categorization of Hispanics and differences among various Hispanic groups. In addition, the small number of Hispanic students in many sources of data leads to instability and unreliability in estimates of trends within that group--a problem that is exacerbated when the scores of Hispanic students are reported separately for different Hispanic groups, such as Mexican Americans and Puerto Ricans. 13/ Given that unreliability, consistency of the trends among a variety of tests is particularly important.

Of the five data sources used in this report that provided trend data on Hispanic students, all but one showed a clear narrowing of the gap between nonminority students and at least one Hispanic group. The sole exception is local data from the Montgomery County (Maryland), public schools, which showed slight and not entirely consistent increases in the size of the gap. 14/

13. Average scores of various Hispanic subgroups could be pooled, but the differences in both achievement levels and recent trends among these groups--documented in this Appendix--argue against that approach when separate tabulations are available.

14. Montgomery County (Maryland) Public Schools, "MCPS Test Results by Racial/Ethnic Groups, 1977-1982" (unpublished, 1982).

The SAT

College Board data distinguish between two Hispanic groups: Mexican Americans and Puerto Ricans.

The narrowing of the gap between Mexican-American and nonminority students has been fairly consistent since the first year of data and appears on both scales (see Figure E-1). Over the full nine years of data, the convergence of scores between Mexican-American and nonminority students is 75 percent or 80 percent as great as that between blacks and non-minority students. As in the case of blacks, the convergence was a bit greater on the mathematics scale than on the verbal scale. The trend among Mexican Americans also parallels that among blacks, in that the relative gains appear to have ended or tapered off in the past few years. The year-to-year fluctuations in the Mexican-American students' scores, however, call this short-term pattern into question.

Puerto Rican students also showed gains relative to nonminority students, but in this case, the gains were both small and far less consistent from year to year, perhaps partly because of the relatively small number of Puerto Rican students taking the SAT (see Figure E-1). The relative gains of Puerto Rican students parallel those of blacks and Mexican Americans in being greater in mathematics than on the verbal scale. On both scales, however, their relative gains were only about 40 percent as large as those of black students over the full nine years.

The NLS and HSB

The NLS/HSB comparison shows relative gains among both Mexican-American and other Hispanic students in all three subjects tested (reading, vocabulary, and mathematics), with Mexican-American students showing a larger relative gain in vocabulary (see Table E-4). With the exception of the vocabulary gains by Mexican Americans, the relative gains of Hispanics were much smaller than those of black students. All of these patterns, however, are open to question, because the Hispanic sample sizes are small. For that reason, even fairly striking changes are not significantly different--in a statistical sense--from no change.

The National Assessment of Educational Progress

The NAEP data show an entirely consistent pattern of relative gains by Hispanic students (not further separated into subgroups) in both reading and

TABLE E-4. AVERAGE ACHIEVEMENT OF HISPANIC AND OTHER STUDENTS IN THE NLS AND HSB, BY SUBJECT

Group	1972	1980	Change
	Vocabulary		
Mexican American	3.47	3.50	0.03
Other Hispanic	4.36	3.71	-0.65
Other a/	7.04	6.22	-0.82b/
	Reading		
Mexican American	6.28	5.60	-0.69
Other Hispanic	6.49	5.72	-0.77
Other a/	10.51	9.57	-0.94b/
	Mathematics		
Mexican American	8.02	7.54	-0.48
Other Hispanic	7.48	7.90	-0.41
Other a/	13.90	12.97	-0.93b/

SOURCE: Rock and others, *Factors Associated with Decline of Test Scores*, Tables D-1, D-2, and D-3.

NOTE: Components might not sum to totals because of rounding.

a. "Other" category includes non-Hispanic whites, Asian Americans, and American Indians.

b. Statistically significant at the .05 level or less.

mathematics--the only subjects for which such comparisons have been made available (see Tables E-5 and E-6). These relative gains are apparent in all three age groups and during periods of both increasing and decreasing scores. They are generally, but not in every case, smaller than those of black students. 15/

The Texas State Assessment

The data from the Texas assessment of mathematics and reading achievement of ninth-grade students is consistent with the other data reported here. Hispanic students on average scored between black and nonminority students, although closer to black students. Moreover, like black students, they gained relative to the nonminority average (see Figure E-4).

[15] Note that in reading, the relevant comparison is the change in blacks' scores from 1974 to 1983, not the change from 1970 that is tabulated in Table E-2. Scores for Hispanics are not available from the 1970 assessment.

TABLE E-5. MATHEMATICS PERFORMANCE OF NONMINORITY AND HISPANIC STUDENTS IN THE NATIONAL ASSESSMENTS
(Average percentage of items answered correctly) a/

	1972 (Estimated) b/	1977	1981	Change 1972-1981
Age 9				
Nonminority a/	60.1	58.1	58.8	-1.28
Hispanic	46.1	46.6	47.7	1.65
Age 13				
Nonminority a/	62.3	59.9	63.1	0.84
Hispanic	48.4	45.4	51.9	3.52
Age 17				
Nonminority a/	66.7	63.2	63.1	-3.56
Hispanic	50.8	48.5	49.4	-1.42

SOURCE: CBO calculations based on National Assessment of Educational Progress, *The Third National Mathematics Assessment: Results, Trends, and Issues*, Table 5.1; and *Mathematical Technical Report: Summary Volume* Tables 2, 3, and 4.

a. Nonminority is non-Hispanic white, labeled "white" in the cited sources.

b. These estimates for 1972 differ from published NAEP results for the 1972 assessment. The published results for that year are based either on the 1972 item pool or on the items used in both 1972 and 1977, while the trend results comparing the 1977 and 1981 assessments reflect items used in both the 1977 and 1981 assessments. In order to circumvent the large disparities in the item sets, 1972 results were estimated here by adjusting the 1977 results (on the items used in 1977 and 1981) by the 1972-to-1977 change (on the items used in 1972 and 1977).

TABLE E-6. READING PERFORMANCE OF NONMINORITY AND HISPANIC STUDENTS IN THE NATIONAL ASSESSMENTS
(Average proficiency scores)

Group	1974	1979	1983	Change 1974-1983
Age 9				
Nonminority a/	215.9	219.7	220.1	4.2
Hispanic	182.9	189.1	193.0	10.1
Age 13				
Nonminority a/	260.9	263.1	263.4	2.5
Hispanic	231.1	236.0	239.2	8.1
Age 17				
Nonminority a/	290.7	291.0	294.6	3.9
Hispanic	254.7	261.7	268.7	14.0

SOURCE: National Assessment of Educational Progress: *The Reading Report Card: Progress Toward Excellence in our Schools*, Data Appendix.

a. Nonminority is non-Hispanic white, labeled "white" in the cited source.