

## 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.<sup>1/</sup> 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.<sup>2/</sup>

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, this tendency has taken several forms, perhaps the most salient of which is 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 shows that some patterns are reasonably consistent among tests and therefore 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

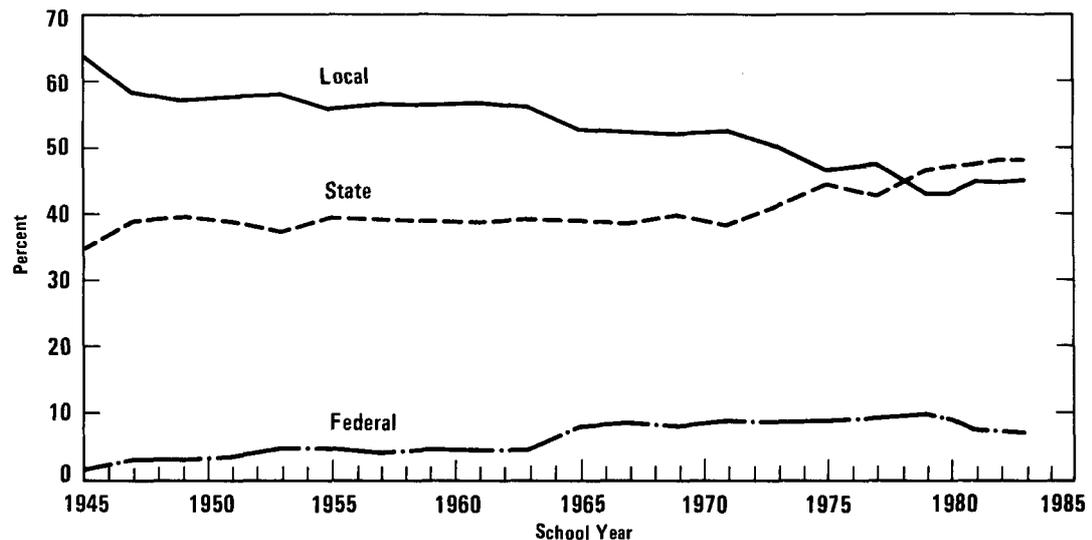
of the money spent for this purpose by all levels of government--education is a truly national concern. Debate about educational policy thus often emphasizes questions of national interest. For example, although there is surprisingly little evidence about the specific skills and abilities that contribute 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.<sup>3/</sup> 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.<sup>4/</sup> 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 the time of the First World War.<sup>5/</sup>

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. <sup>6/</sup> 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.<sup>7/</sup> 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.<sup>8/</sup>

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.<sup>9/</sup>

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.<sup>10/</sup> 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.<sup>11/</sup> 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.<sup>12/</sup> 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.<sup>13/</sup> 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. 393.

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.<sup>21/</sup> The Administration has emphasized its information-dissemination role in attempts to prompt reforms.<sup>22/</sup> 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.<sup>23/</sup>

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.

1

1

## 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, while the limitations of standardized tests are less obvious, they can be severe. <sup>1/</sup>

Perhaps the most important strength of standardized tests is that they can be freed of much of the subjectivity that can plague such 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.<sup>2/</sup> 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.<sup>3/</sup> 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:

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

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

- o Ascertain whether students have acquired specific skills or information;
- o Rank students in terms of their knowledge or skills; or
- o 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. <sup>5/</sup>

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.

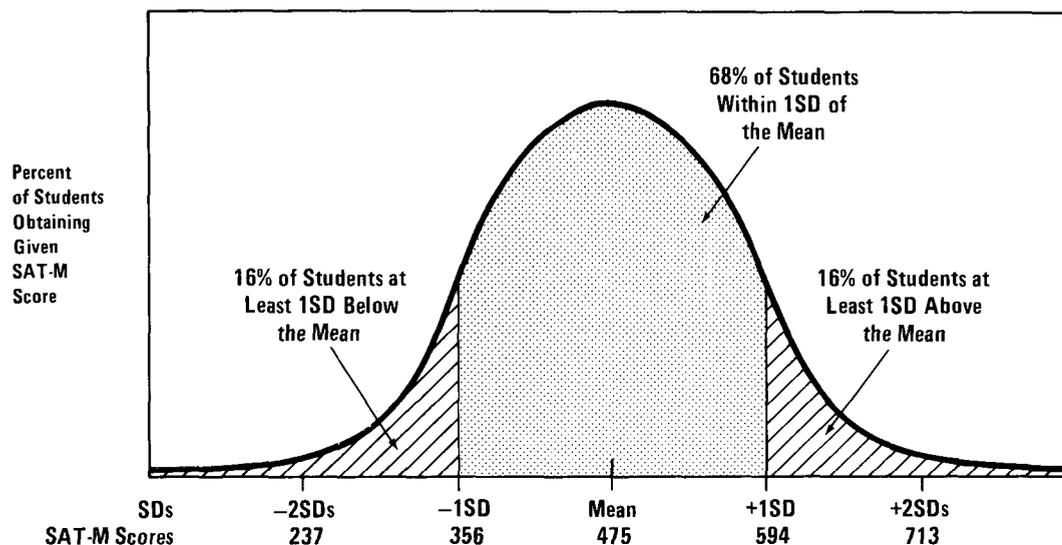
Norm-referenced tests are often relatively free of the floor and ceiling effects that can plague interpretation of MCTs. Since norm-referenced tests are designed to rank students, they typically must be easy enough to differentiate among low-achieving students but difficult enough to discriminate at 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).<sup>6/</sup> 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.<sup>7/</sup> 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.<sup>8/</sup>

#### 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 level 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.