DOCUMENT RESUME

ED 411 290 TM 027 298

| | |
|---|---|
| TITLE | Measuring Student Achievement. |
| INSTITUTION | Ohio State Legislative Office of Education Oversight, Columbus. |
| PUB DATE | 1993-09-00 |
| NOTE | 17p. |
| PUB TYPE | Reports - Evaluative (142) |
| EDRS PRICE | MF01/PC01 Plus Postage. |
| DESCRIPTORS | *Academic Achievement; *Achievement Tests; Criterion Referenced Tests; Elementary Secondary Education; Intelligence Tests; Measurement Techniques; Multiple Choice Tests; Norm Referenced Tests; Outcomes of Education; *Performance Based Assessment; *Portfolio Assessment; Portfolios (Background Materials); Standards; *State Programs; Test Use; *Testing Programs |
| IDENTIFIERS | *Ohio |

ABSTRACT

This report provides background on statewide student testing and testing alternatives in Ohio. The status of achievement testing in other states is also discussed. Educational testing historically has included the two strands of intelligence testing and achievement testing. This report focuses on achievement testing. The two basic forms are criterion-referenced tests, which compare student achievement to a set of desired outcomes or standards, and norm-referenced tests, which compare achievement to that of other students who took the same test. Three common methods of statewide testing include written multiple-choice tests, performance-based tests, and portfolio assessment. To be successful at performance-based tests, students must be able to demonstrate their understanding of a subject through complex or multistage problems that require the application of several skills at once. Portfolio assessment requires that samples of student work be collected over a period of time. To prepare for portfolio assessments, teachers provide feedback on how work can be improved and help students understand the criteria for selecting their best work for inclusion in a portfolio. How best to test students on a statewide basis depends on how the tests are to be used. Testing requirements differ among states, but 47 states require districts to test public school students at some time between grades 1 and 12. There is a national trend toward the use of performance-based tests or portfolio assessments. In Ohio, the ninth-grade proficiency test consists of three criterion-referenced multiple-choice tests (mathematics, reading, and citizenship) and one criterion-referenced performance-based test (writing). An appendix discusses the early history of testing and the use of intelligence tests. (Contains seven exhibits.) (SLD)


# MEASURING STUDENT ACHIEVEMENT



*LEGISLATIVE OFFICE OF EDUCATION OVERSIGHT*
*COLUMBUS, OHIO*
*September 1993*

*LEGISLATIVE OFFICE OF EDUCATION OVERSIGHT*
*30 EAST BROAD STREET - 27TH FLOOR*
*COLUMBUS, OH 43266*
*Telephone: (614) 752-9686*



The Legislative Office of Education Oversight (LOEO) serves as staff to the Legislative Committee on Education Oversight. Created by the Ohio General Assembly in 1989, the Office studies of education-related activities funded wholly or in part by the state of Ohio.

This is a report of the LOEO to the Legislative Committee on Education Oversight. ***This report of the LOEO staff does not necessarily reflect the views of the Committee or any of its members.***

September, 1993

# MEASURING STUDENT ACHIEVEMENT

This report provides background and context on statewide student testing and testing alternatives. Information about achievement testing in general and some of the differing perspectives on the purposes and uses for standardized testing and test results is included. The status of achievement testing in other states is also provided.

## INTRODUCTION

The United States has a long history of student testing. Most education experts agree that "testing has helped shape the form and substance of American education" (Corbett & Wilson, 1991). Student testing and reforms in education often are considered simultaneously.

Educational testing historically has included two separate but related strands: intelligence testing and achievement testing. This report focuses on achievement testing; intelligence testing is discussed in the Appendix. Tests have played a prominent role in the practice of sorting and tracking students for purposes of vocational selection and preparation, and determination of college readiness. It is not clear whether these practices have led to significant improvements in instruction or student learning.

## FORMS OF ACHIEVEMENT TESTING: CRITERION- AND NORM-REFERENCED

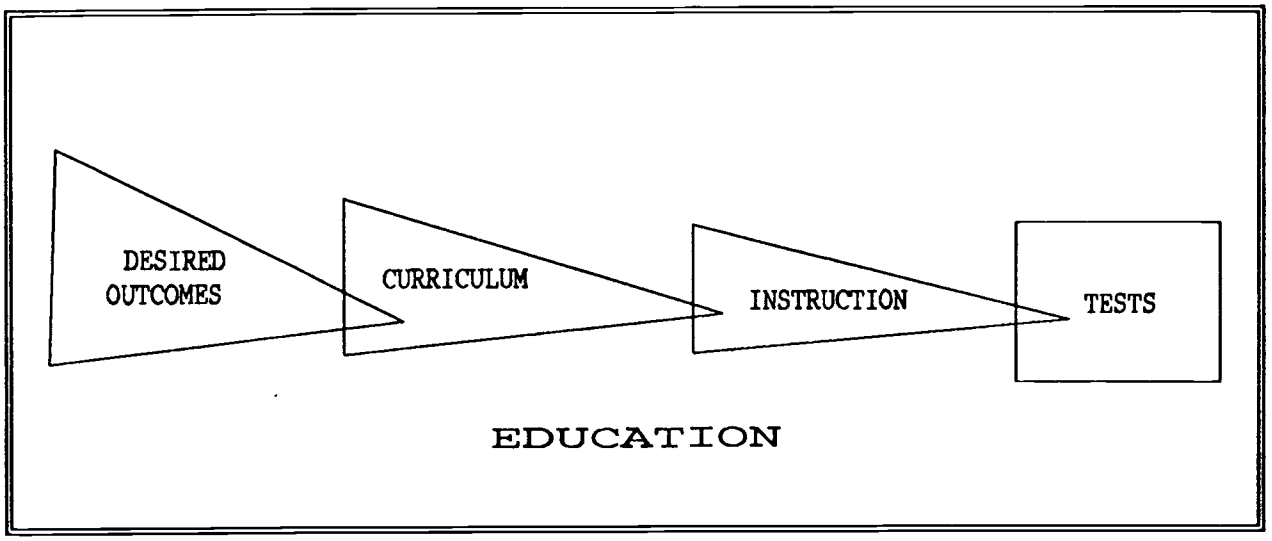
Achievement tests measure how much students have learned in specific subject areas. There are two basic forms of achievement tests: criterion-referenced tests and norm-referenced tests. Criterion-referenced tests compare students' achievement to a set of desired outcomes or standards, while norm-referenced tests compare students' achievement to that of other students who took the same test at a previous time.

Statewide tests are standardized tests. Standardization means that every test is administered in the same manner, for the same purpose, and according to prescribed directions; standardized tests are scored and interpreted in conformance with defined rules. Standardized tests can be developed by independent, commercial businesses or by the public education authority.

Test development requires the application of rigorous statistical procedures to ensure that the tests accurately and reliably measure student achievement. Although the technical aspects of test development are important, decisions about what to test and how to use test results are equally important.

Criterion-referenced achievement tests measure students' progress relative to predetermined educational outcomes or standards. Ohio's ninth-grade proficiency test is an example of a criterion-referenced test. Outcomes define the subject matter a student is expected to know, and the tasks a student should be able to perform. As such, the desired outcomes are translated into what is taught to students, and what is taught, ideally, becomes what is tested. Exhibit 1 demonstrates the relationship between desired outcomes, curriculum, instruction, and tests.

EXHIBIT 1
RELATIONSHIP OF OUTCOMES TO TESTS




Questions or items on criterion-referenced tests are taken directly from the curriculum or educational material appropriate for the grade level(s) of students being tested. Test scores, therefore, are intended to indicate the extent to which the material has been mastered.

Norm-referenced achievement tests are designed to measure achievement by comparing an individual student's performance with that of others. Test scores do not correspond to a particular set of educational outcomes. Instead, these scores describe the relative rank or position of a student compared to a group of students who have taken the test at some previous time.

Questions on the norm-referenced test are developed from generally defined subject areas which may or may not correspond to students' current curricula. In particular, items on the test are selected or replaced based on their ability to differentiate or separate students. A norm-referenced test is designed so that 50% of the students score above average and 50% score below average.

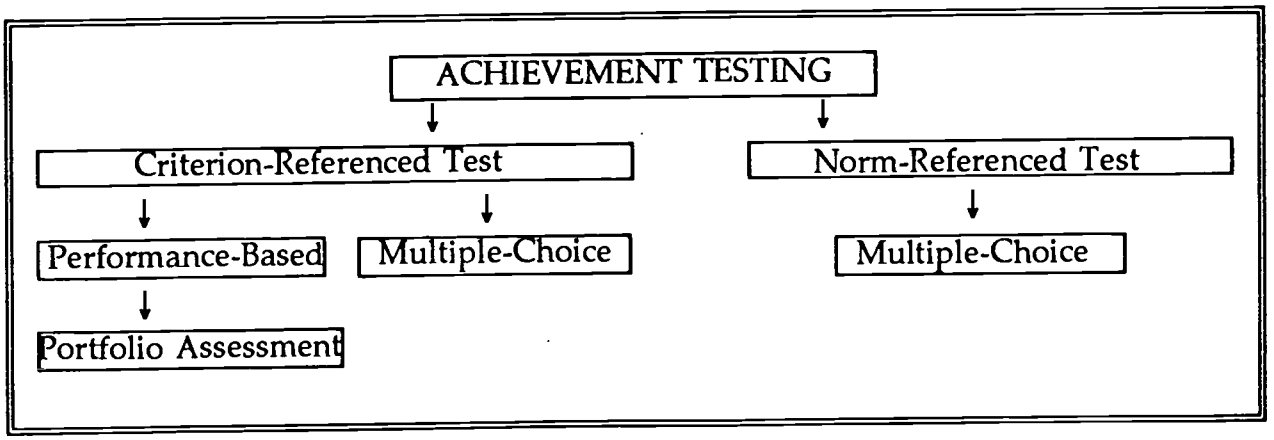
Norm-referenced test publishers do not compute a yearly national average by which currently tested students can be compared with all other currently tested students. Instead, the publishers identify a "norming group," representative of students across the country. These students are given the test when it is first developed. Students who later take the test are compared against the scores of this norming group.

As students may take the same version of the test several times, they and their teachers become more familiar with the test contents. This situation has led to the criticism that the scores of the currently tested students may be artificially higher than those of the original norming group (i.e., "all the children are above average").

## TESTING METHODS: MULTIPLE CHOICE, PERFORMANCE, AND PORTFOLIO

There are several methods by which statewide student testing can occur. Each method, however, embodies an implicit theory of how learning occurs, and as such, carries implications for how instruction will occur. A test is most useful when it is selected with a full understanding of the education and learning theory on which it is based. Three common methods of statewide testing include: written multiple-choice, performance-based, and portfolio assessment. Exhibit 2 depicts how the two forms of achievement testing (criterion-referenced and norm-referenced) are related to three testing methods.

**EXHIBIT 2**
**FORMS OF ACHIEVEMENT TESTING AND METHODS**



Multiple-choice tests

To be successful at taking multiple-choice tests, students must be skilled at recognizing correct answers. Test questions are often based on isolated skills and unrelated facts. An answer is selected from a given list by recalling facts, making mathematical computations, or applying information that is provided within the context of a test question. Exhibit 3 contains examples of questions from standardized, multiple-choice tests for primary grades.

To prepare students for multiple-choice tests, teachers help them learn how to choose correct answers. Many teachers use worksheets that emphasize drill and repetition to help their students memorize facts and practice skills. As a result, the students may be able to recognize many facts about a given topic, yet have limited understanding of it. For example, a student may be able to choose from short lists the correct names of clouds, indicate that some thermometers use mercury to show temperature, and pick the correct freezing point of water from a series of numbers. However, the student might have no understanding of how weather is made--why rain falls, how clouds form, or why seasons cause changes in temperature.

EXHIBIT 3

SAMPLE MULTIPLE-CHOICE TEST QUESTIONS

| | | |
|---|---|---|
| MATH | 7<br>+ 3 | a. 4<br>b. 10<br>c. 37<br>d. 73<br>e. none of the above |
| SCIENCE | Which of these has bark?<br>A. fish<br>B. tree<br>C. flower<br>D. crocodile | |
| READING | Bill picked up gloves from the desk. He took his jacket out of the closet. At the back door he put on his boots. He walked to the garage and took out the snow shovel.<br>What did Bill take from the closet?<br>A. boots<br>B. jacket<br>C. shovel<br>D. gloves | |

Performance-based tests

To be successful at performance-based tests, students must be able to demonstrate their understanding of a subject. Test questions on performance-based tests pose complex or multi-stage problems that require students to apply several skills at once and to integrate information from various sources. Exhibit 4 contains examples of questions from standardized, performance-based tests of other states.

To prepare students for these tests, teachers design classroom activities that help students understand the relationships among pieces of information. Teachers encourage students to explore diverse ways to solve problems and answer questions. Students are taught to use information, not just recognize it. Teachers use projects, oral presentations, essays, and experiments to improve students' abilities to think critically and communicate persuasively.

It is unlikely that a student would successfully complete tasks on performance-based tests without a thorough understanding of the subjects being tested. For example, a performance-based test might ask a student to discuss why certain kinds of clouds are more prevalent during a particular season, and what conditions affect the formation of those clouds. Not only must the student recognize the names of the specified clouds, but he must understand how and why clouds exist.

Portfolio assessment

Portfolio assessment requires that samples of each student's work be collected over a period of time. Portfolios are a type of performance-based test because they require that students demonstrate their competencies, instead of choosing an answer from a given list. But rather than elicit responses from students during a single, structured situation, portfolios collect samples of students' classroom work over a period of months or years. To prepare for portfolio assessments, teachers provide feedback on how work can be improved and help students understand the criteria for selecting their best work for inclusion in a portfolio.

Portfolio assessments are standardized to the extent that there are prescribed guidelines, by subject area, for what must be included in a portfolio, and prescribed criteria for how the portfolio material will be scored. Standards for portfolio assessments and other methods of testing can be set by educators and policy makers. Exhibit 5 is an example of guidelines for a math portfolio.

EXHIBIT 4

## SAMPLE PERFORMANCE-BASED TEST QUESTIONS

High School Math Exam

Two banks are offering car loans with monthly payments of $100. One has an interest rate of 16%. The other has a higher rate of 18% but also offers a television worth $400 to those taking out loans. If you need a $6,000 loan and would also really like a color TV, which bank should you use? Explain your answer.

Fourth Grade Science Test Station #1

Measuring Objects - Use the equipment at station #1 to answer the questions below.

| | |
|---|---|
| Equipment: | double pan balance, pile of pennies, jar of water, glass, measuring cup, ruler, thermometer |
| Questions: | How many pennies heavy is the empty glass?<br>How tall is the glass?<br>How much water is needed to fill the glass to the line?<br>What is the temperature of the water? |

Heritage High School English Exam

Read the (provided) passage. In an organized and detailed essay, summarize its main ideas and then explain why you would agree or disagree with what the article says. Support your view with specific examples.

In general, American Society is strongly focused on respecting the rights of others. Some feel that this view of equality should also apply to animals, both domestic and wild. List some responses to the following questions.

a. Do you believe that animals should be used in medical experiments? Agree? Disagree? Why?

b. Do you believe that zoos are necessary to save our wild animals? Agree? Disagree? Why?

c. Do you believe that animals raised for domestic consumption have rights and should be treated humanely? Agree? Disagree? Why?

EXHIBIT 5

## SAMPLE PORTFOLIO GUIDELINES

California Math - Portfolio Guidelines

a. Portfolios should contain samples of:
   (i) "typical" work (as judged by the teacher)
   (ii) several "best efforts" as judged by the student in consultation with the teacher
   (iii) student reflections as to why each piece was selected.

b. Guidelines for works included in the portfolio:
   (i) four pieces of individual work with interesting and challenging mathematics problems. Choose two from the beginning and two from the end of the course. Student writing must accompany work.

   (ii) one piece representing a small-group investigation. The written report should include: problem, difficulties encountered, final conclusions or product. (Group report is permissible.)

   (iii) one-two individual reflective or imaginative piece(s).

   (iv) one-two other pieces revealing strength of program, evidence of persistence, etc.

c. Some possible reflective topics:
   (i) What are the most important ideas in mathematics you have learned this year?

   (ii) You have been given newspaper articles with conflicting data. Read the articles carefully and write an analysis of what you think are the facts of the case.

   (iii) If you lived in "Flatland" (in which there are only two spatial dimensions), how would the three objects on the desk look to you?

## PURPOSES AND USES FOR TESTING

How best to test students on a statewide basis and which method(s) to select depend on how the results are to be used. According to the Educational Testing Service, statewide testing programs are generally used for accountability, that is, monitoring school or district performance; grade promotion or high school graduation; identification of students who need special assistance; and, for the distribution of funds. Approximately one third of states require students to pass a test in order to graduate from high school.

Norm-referenced tests have been developed by educators to sort students into special programs and into curriculum tracks. Such tests tell educators where a particular student scores in relation to a "normal curve." For example, in order to identify children to be placed in remedial classes or classes for gifted and talented students, it is necessary to know whether a child is achieving more or less than his peers. Likewise, the practice of sorting students into academic or vocational tracks is associated with the development of norm-referenced tests (see Appendix).

Criterion-referenced tests, on the other hand, are useful for determining whether a student has mastered the skills and knowledge of a particular curriculum, regardless of whether she does better or worse than others taking the test. These tests are used to help individual students focus on specific skills not yet mastered. Teachers can use the tests to assess how well students are learning the intended knowledge and skills. Results from criterion-referenced tests are linked to changes in instruction and curriculum.

Performance-based tests and portfolio assessments require greater expenditures of staff resources and student time than multiple-choice tests. However, because they are more direct measures of achievement they can be incorporated into learning tasks. Multiple-choice tests are efficient to administer and easy to grade, but are separate from classroom activities and not an integral part of learning. Nationally, it is estimated that the administration of statewide, multiple-choice achievement tests takes as much as eight school days per year.

Given the cost and nature of performance testing and portfolio assessment, testing every child may be seen as prohibitive. However, testing a statewide sample of students using these methods can yield information on how the state as a whole is doing on desired learning outcomes. The choice of testing all students or a sample of students depends on the specific purpose of the test.

Statewide multiple-choice tests are limited in the type of learning they can assess. Critical thinking, reasoning, writing, understanding, problem solving, discerning, decision making, and creativity are examples of traits not typically measured by multiple-choice achievement tests. They provide no information about why students choose right or wrong answers and no information about the thought process used to choose their answers.

## STUDENT TESTING IN OTHER STATES

Testing requirements differ among states. Implementation time frames, types of tests, number of students tested, purposes for mandated testing, and linkage with other aspects of education reform cover a wide spectrum.

A 1990 report issued by the Educational Testing Service indicated that 47 states required local school districts to test public school students at some point between grades one and twelve. Although this represented an increase of only five states from the previous five-year period, many states broadened their testing programs during that same time:

- eleven states added new grade levels to be tested, including pre-kindergarten and pre-first grade;
- six added science and social studies to their testing program, and others added writing, especially essays, to replace multiple-choice exams in language arts;
- two states moved from testing representative samples of students in a grade to testing all students; and
- three states switched from allowing local school districts to choose their tests to mandating the use of a state-selected instrument.

There is a national trend toward replacing or supplementing statewide multiple-choice tests with alternatives, such as performance-based tests or portfolio assessments. This trend reflects the efforts of states to engage students in more rigorous learning.

According to the annual survey by the Association of State Assessment Programs, in 1991, 68% of the 38 reporting states used alternative methods or were trying out alternative methods in their statewide testing programs. In 1992, 72% of 50 reporting states indicated the use of alternative methods alone or in combination with multiple-choice testing to assess achievement across academic subjects.

Thirty-two percent of those reporting in 1991 indicated that they relied solely on criterion- or norm-referenced, multiple-choice tests. In 1992, only 18% relied solely on this method. In each year, only two of the reporting states used alternative methods exclusively.

Twenty of the 26 states using alternative assessment methods in 1991 applied the methods in both elementary and secondary schools. Four of the states used alternative testing methods in elementary schools only.

Survey results of 1991 also indicated that states most often use an alternative method to assess writing skills. Of the 38 reporting, 20 states assessed writing skills of students with standardized, performance-based tests such as written essays and writing portfolios. As shown in Exhibit 6, reading and math skills were the next most often tested subjects using alternative assessment methods, primarily through performance-based testing and open-ended test items. Ohio uses an alternative method to test students' writing achievement in the ninth-grade proficiency test.

EXHIBIT 6
FREQUENCY OF ALTERNATIVE TEST METHODS BY SUBJECT
(Non-Multiple-Choice)

| SUBJECT AREA | NUMBER OF STATES* |
|---|---|
| Writing | 20 |
| Reading | 15 |
| Math | 15 |
| Science | 9 |
| Social Studies | 8 |
| Other | 5 |

* 38 states responded to the 1991 survey by the Association of State Assessment Programs

The 1991 survey indicated that 25 of the 38 reporting states administer statewide achievement tests on an every-student basis. Four states sample students to be tested. Another eight states combine the practice of every-student testing with sampling test

questions from a pool of possible questions. In one state, testing is under development. In 1991, 14 of the 38 states required tests for high school graduation; in 1992, 18 of the 50 states required such tests (see Exhibit 7).

## EXHIBIT 7
## FREQUENCY OF SAMPLING STUDENTS FOR TESTING

| WHO IS TESTED | NUMBER OF STATES* |
|---|---|
| Students sampled | 4 |
| Test item sampled but every student tested | 8 |
| Every student tested | 25 |

* 37 states responded to the 1991 survey by the Association of State Assessment Programs

## SUMMARY

Achievement tests measure how much students have learned in specific subject areas. There are two basic forms of achievement tests. Criterion-referenced tests compare students' achievement to a set of desired outcomes or standards, while norm-referenced tests compare students' achievement to that of other students.

Statewide student tests can be administered through several methods. Performance-based methods require students to demonstrate their achievement. They can be incorporated into learning tasks. Some performance-based tests measure students' achievement at one point in time. Others, like portfolio assessments, compile students' achievement over time. The multiple-choice method is considered to be less costly than performance-based methods, however, it is more limited in the type of learning it can assess.

Ohio's ninth-grade proficiency test consists of three criterion-referenced, multiple-choice tests (mathematics, reading, and citizenship) and one criterion-referenced, performance-based test (writing).

State boards of education and state legislators across the country are assuming greater roles in student testing, including setting statewide standards, determining minimum competencies, and financing testing mandates. Policy makers are becoming

familiar with many of the technical aspects of testing in order to reach informed decisions about which form and method of testing to use, as well as who is to be tested, when, and how often.

States differ widely, both in terms of the kinds of tests mandated and the methods used to test students. Nationally, statewide achievement testing reflects two movements: first, away from norm-referenced tests toward criterion-referenced tests; and second, within criterion-referenced tests, from multiple-choice toward performance-based tests. This trend reflects the efforts of states to engage students in more rigorous learning.

Criticisms of the limited utility of multiple-choice tests and the time and resources required for performance-based tests have led states to develop statewide testing programs that combine the methods. States have added to or replaced multiple-choice tests with ways to assess their students' demonstrated knowledge and skills on more complicated tasks than can be assessed if testing is limited to what can be shown via multiple-choice, paper-and-pencil tests.

The strong relationship of testing to instruction has become apparent in light of discussions regarding education reform. If the test requires students to recognize and choose a response from a predetermined set of answers based on isolated skills and knowledge, then teaching focuses on preparing students for this task. If the test calls for constructing a response to a complex problem in order to demonstrate in-depth understanding of a subject and the integration of several skills, then teaching focuses on helping students to accomplish these tasks. In effect, what is measured is taught; the assessment must match the desired learning.

# APPENDIX
# EARLY HISTORY OF TESTING

## Intelligence testing

Standardized testing of students began in 1904 when Alfred Binet attempted to assess the learning abilities of mentally retarded children in France. His goal was to develop a technique that could be used to identify children who required specialized help.

> Assembling a series of short tasks that were related to everyday problems of life and that required basic processes of reasoning without incorporating learned skills such as reading and computation, he assigned an age level to each task and the relationship between a child's chronological age and his or her mental age became known as the 'intelligence quotient' [IQ]. (Corbett & Wilson, 1991)

Binet cautioned that the IQ number is not an entity in itself nor is it a device for ranking "normal" children. He further cautioned that low scores cannot be used to mark children as innately incapable.

Dr. Binet's intentions were significantly altered when the IQ test was brought to America. In the early 1900s, IQ testing was promoted as a screening device to "curtail breeding . . . of feeble-minded" (Gould, 1981), and "to categorize new immigrants landing on Ellis Island" (Corbett & Wilson, 1991). The test was modified for use on a full range of the population to categorize people into potential occupations. The IQ test was used during World War I by the military as a "quantifiable means for determining the worth and position of different recruits" (Corbett & Wilson, 1991).

The IQ test as it had evolved was administered to post-war college undergraduates to identify those suspected of neglecting their studies. In 1925, the College Board requested an intelligence test for college admission be prepared to test the aptitude of junior and senior high school students for college-level work. The Scholastic Aptitude Test (SAT) was introduced. The volume of testing declined during the depression, but the growth of mental measurement reemerged with the onset of World War II.

## Achievement testing

The development of standardized achievement tests for children began at the turn of the century as an attempt to measure and predict educational success. University psychologists developed subject-matter tests during experiments to explore the

relationship between the amount of time spent on subjects, such as spelling or math, and achievement in that subject area. Through their work "they initiated the idea of testing at all grades, out of which grew the notion of test norms" (Corbett & Wilson, 1991).

The United States was also experiencing great expansions in school enrollment at the turn of the century. In 1870 the secondary school enrollment was only 80,000; by 1910 the enrollment had reached 900,000. A more diversified high school curriculum emerged from the expansion in the student population and subject-matter tests quickly gained popularity as an "objective" method for sorting the increased number of students and standardizing the schooling process. More varied programs and "tracks" were introduced "so that failure would become unnecessary" (Resnick, 1982).

"The prevention of dropouts, which in earlier times had not been a major social concern, became more important with the changing job market requiring more skills, and the large immigrant population needing socialization" (Corbett & Wilson, 1991). Educational testing was perceived as an efficient means to place students into homogeneous groups to the end of keeping students in school over a longer period of time.

Concern for efficiency and accountability of schools grew during the early 1900s. The business/industrial community promoted "centralizing school administration, controlling school costs, and setting standards by which to compare school performance within and across [school] districts." Achievement "testing became an important tool for keeping school systems administratively efficient and locally accountable" (Corbett & Wilson, 1991).

In summary, both norm-referenced achievement testing and intelligence testing have played a prominent role in the practice of sorting and tracking students for purposes of vocational selection and preparation, and determination of college readiness. Test scores also became the criteria for standardization and accountability within schools.

## REFERENCES

Corbett, D. & Wilson, B. (1991). Testing, Reform and Rebellion. Norwood, New Jersey: Ablex Publishing Corporation.

Gould, S.J. (1981). The Mismeasure of Man. New York, New York: Norton.

Resnick, D. (1982). "History of Educational Testing." In A.K. Wigdor and W.R. Garner (Eds.), Ability Testing: Uses, Consequences, and Controversy. Washington, D.C., National Academy Press.


U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)


# NOTICE

## REPRODUCTION BASIS

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☑ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").