SOME HISTORICAL ASPECTS OF TESTING AND THEIR IMPACT ON EDUCATION

A thesis written in partial fulfillment of the requirements for the degree of Master of Arts.

Herbert C. Fillmore
St. Mary's University
School of Education
March, 1965.

TABLE OF CONTENTS

		Page
INTRODUCTION		1
Chapter		
I.	EARLY EDUCATIONAL MEASUREMENT TO THE NINETEENTH CENTURY	3
II.	EDUCATIONAL TESTING EMBRACING THE NINETEENTH CENTURY	7
III.	THE FIRST FIFTY YEARS OF THE TWENTIETH CENTURY	14
IV.	RECENT TRENDS AND FUTURE DEVELOPMENT OF EDUCATIONAL EVALUATION	27
V.	TESTING: AN EDUCATIONAL INNOVATOR	34
BIBLIOGRAPHY		50

INTRODUCTION

Measurement of human behavior with primary reference to the capacities and educational attainments of school children has been divided by Greene, Jorgensen and Gerberich into three time periods.

During the first period, from the beginning of historical records down to about the nineteenth century A.D., educational measurements were naturally quite crude. Although the fact that individual differences has been recognized for several thousand years and educational measurement made formal entrance to the schools as early as medieval times, relatively little progress in educational testing was made until the nineteenth century. During the second period, embracing approximately the nineteenth century, educational measurement began to assimilate from various sources the ideas and the scientific techniques which were later to result in the modern objective testing movement. The brief third period, dating from nineteen hundred to the present, has been characterized by tremendous advances in statistical techniques, in the measurement and evaluation of achievement, intelligence and personality, and in the classroom use of test results.

The purpose of this thesis is to present a history of testing by reviewing the literature. Some implications for the future of testing are given, not as a warning, but rather in the hope that educators will be flexible enough to make evaluation an integral part of the educative process.

Finally, with apologies to the curriculum developers who too long have held the notion that curricula modify education, I submit that the testing movement is the prime innovator of education.

This thesis is founded on the premise that the object of education is man himself - his learning, aptitudes, interests, ability and personality. The worthiness of his educational system is reflected in his behavior as it is modified in the light of his learning. Measurement is a technique, a part of the evaluative technology, which has and can continue to sharpen our inquiries about ourselves; to enlarge our vision of man's potential and achievement; to deepen our humility. Without it we will fail to improve our practices for we will not know whether they are effective or ineffective.

CHAPTER I

EARLY EDUCATIONAL MEASUREMENT TO THE NINETEENTH CENTURY

The first evidence of the oral examination is found in the Old Testament. It concerns the test the Gileadites gave to the enemy Ephraimites who wished to cross the Jordan. When asked to pronounce the word "Shibboleth" the Ephraimites could only respond "Sibboleth", whereas friendly tribes responded with the correct pronounciation. Forty-two thousand Ephraimites were killed because they failed to pass this objective test.[1] Socrates, in a method made famous, subjected his pupils to exhaustive and searching questioning. Oral quizzing, Socratic or otherwise, has undoubtedly been a part of classroom procedure from the beginning of teaching activity - in fact, there have been and still are times when, for certain teachers, it constitutes practically the whole of the teaching act.

Written tests are probably of more recent

---

[1] Old Testament, Holy Bible (Cleveland: The World Publishing Co., 1947), Judges 12: 5-7.

origin than oral quizzes, but even they date back many centuries. As early as 2000 B.C., China had a national system of civil service examinations. These exams have been known down through the ages for their unusual severity. Confined in isolated cells, candidates were compelled to write lengthy dissertations on assigned topics.[1]

Individual differences among people have long been recognized. Plato, nearly four centuries B.C., divided his ideal society into three classes, workers, protectors and rulers. He believed that persons suited to each class should receive education for the fullest development of their personalities.[2] Quintillian, shortly after the start of the Christian era, wrote that masters should observe differences in ability and inclinations of persons they instruct for the "forms of the mind are not less varied than those of bodies."[3]

The first use of tests for the measurement of the results or outcomes of education were probably

---

[1]W. A. P. Martin, The Chinese: Their Education, Philosophy, and Letters (New York: Harper and Brothers, 1881), pp. 45-49.

[2]Edgar W. Knight, Twenty Centuries of Education (Boston: Ginn and Co., 1940), p. 62.

[3]William Boyd, The History of Western Education (London: A. and C. Black Ltd., 1921), p. 76.

not unlike certain of the performance tests of today, at least to the extent that they measured physical performance and that they were not paper and pencil tests.

Among various primitive tribes, in which the young men were taught to hunt, fish and fight, the initiation ceremonies prerequisite to their admission to the ranks of adult males tested knowledge of tribal customs, endurance, bravery and other skills and abilities thought necessary for tribal protection.[1]

The ancient Spartans, whose education curricula for their youth stressed physical development and stoicism, conducted examinations as early as 500 B.C. in which the young men underwent painful ordeals.[2] In ancient Athens, the stress upon athletics and aesthetic development led to evaluation by means of games and contests and of reading, writing, and singing ability.[3]

In medieval times, the oral examination was used in universities. The University of Bologna

---

[1]Charles Russell, Standart Tests (Boston: Ginn and Co., 1930), pp. 14-15.

[2]Ibid, p. 16.

[3]Knight, Twenty Centuries. . . . , pp. 52-53.

by 1219 A.D., and the University of Paris before the close of the thirteenth century, required degree candidates to defend their theses orally. However, the written educational examination probably made its first appearance for educational use at Cambridge, England, in 1702.[1]

---

[1]Albert R. Lang, Modern Methods in Written Examinations (Boston: Houghton Mifflin Co., 1930), pp. 2-3.

CHAPTER II

EDUCATIONAL TESTING EMBRACING THE NINETEENTH CENTURY

The first examinations of note in the United States were written in Boston in 1845.[1] Prior to that date, the school committee had orally examined all Boston pupils, or at least those in the highest class in each school, annually. As the pupils increased in numbers, this task became onerous and eventually received only perfunctory attention. Finally, the subcommittee appointed to survey the grammar departments of the schools in 1845, decided to use written examinations in lieu of the time consuming oral examinations, in the fields of arithmetic, astronomy, geography, grammar, history and natural philosophy. These tests were used to rank the schools in order of merit.

This Boston examination project is truly a highlight in the history of education in the United States. It made a great impression on Horace Mann, who at that time was secretary of the Massachusetts Board of Education. It is difficult to ascertain his influence on

---

[1] Otis W. Caldwell and Stuart A. Courtis, Then and Now in Education, 1845-1923 (New York: World Book Co., 1923), i-iii.

the sub-committee, but, as its secretary, his influence was probably considerable and the examinations were some reflection of his ideas. As editor of the Common School Journal, he published extracts from the report and made noteworthy comments on the subject of examinations.[1]

He concluded that the new written examination was so superior to the old oral quiz that no school committee would ever lapse into the former inadequate and uncertain practice. Mann further advanced several reasons to support written examinations.[2] It is interesting to note that although Mann's ideas are much the same as those represented by modern tests, the instruments were however inadequate. In successive issues of the Common School Journal, Mann suggested most of the elements that are found in the modern measurement and evaluation movement.

To the Reverend George Fisher, an English schoolmaster, goes the credit for devising and using what were probably the first objective measures of achievement. His "scale books", used in the Green-

---

[1]Ibid, pp. 237-272.

[2]Ibid, p. 37.

wich Hospital School as early as 1864, provided means for evaluating accomplishments in handwriting, spelling, mathematics, navigation, Scripture knowledge, grammar and composition, French, general history, drawing and practical science.[1] In such subjects as handwriting and drawing, where qualitative rather than quantitative evaluation was the custom, specimens of pupils' work were compared with "standard specimens" to determine numerical ratings. The numerical values for spelling and other subjects to which quantitative measures of achievement were commonly applied depended upon errors in performance.

Although Fisher's "scale books" included the germ of many of the ideas that are incorporated in our present-day educational scales, his work produced no lasting results because he lived too far in advance of the thought and educational practice of his day.

In America, the real inventor of the comparative test was Dr. J. M. Rice, who, in 1894,[2] hit

---

[1]E. B. Chadwick, Statistics of Educational Results (The Museum, A Quarterly Magazine of Education, Literature and Science, Jan. 1864), iii, pp. 480-484.

[2]Leonard P. Ayres, History and Present Status of Educational Measurements (Bloomington: Public School Publishing Co., 1918), p. 11.

upon the idea he developed so effectively that it became the foundation of objective measurement in education. Rice, having administered a list of spelling words to pupils in many school systems and having analyzed the results, confounded the educators at the 1897 session of the Department of Superintendence of the National Education Association with the declaration that pupils who had studied spelling thirty minutes a day for eight years were no better spellers than children who had studied the subject fifteen minutes a day for eight years. Rice was attacked and reviled for this "heresy", and some educators even attacked the use of a measure of how well pupils could spell for evaluating the efficiency of spelling instruction. They contended that spelling was taught to develop the pupil's minds and not to teach themmto spell. It was more than ten years later that Rice's pioneering resulted in significant attention to the objective method in educational testing.[1]

It was not, apparently, until 1796, that individual differences in mental abilities were first brought not under the microscope, but, literally, the telescope. It was in that year at the Greenwich

---

[1]Ibid, p. 12

Astronomical Observatory in England that one of the observers who recorded the instant of time at which stars crossed the lines on telescope lenses was discharged because his observations consistently differed slightly from those of his colleagues. In 1816, however, it was discovered by an astronomer who read an account of this incident that an error of observation, called the "personal equation", characterized the work of all observers and that the amount of error varied from person to person and also in the same person from time to time.[1]

Galton, with the publication of his Hereditary Genius in 1869, brought the scientific study of individual differences into focus and developed it further by instituting measurement of various human physical traits and motor abilities, and even investigated mental ability by methods which many years later became highly fruitful.[2]

Galton's most important contribution to educational measurement was not in the field of individual differences, however, but in the derivation of statistical methods. Here, in devising a system of "stand-

---

[1]Anne Anastasi and John P. Foley, Jr., Differential Psychology (New York: Macmillan Co., 1949), pp. 7-8.

[2]Joseph Peterson, Early Conceptions and Tests of Intelligence (New York: World Book Co., 1925), pp. 73-75.

ard scores" and in developing graphically the idea for objective measure of relationship, the correlation coefficient, he furnished tools essential not only to the development of educational and mental testing, but also to scientific method in education. Pearson later formulated the method now most commonly used for calculating the correlation coefficient.[1]

Dr. E. S. Chaille, an American physician, is credited as early as 1887 with the development of standards and simple tests for judging the mental levels of children to the age of three and with having implied, although not definitely used, the concept of mental age as an index of mental maturity.[2]

Cattell apparently first used the term "mental test" in 1890, almost at the beginning of the period during which scientific method was first applied to the measurement of mental ability.[3] Attempts during the last decade of the nineteenth century by Cattell

---

[1]Henry E. Garrett, Great Experiments in Psychology (New York: Appleton-Century-Crofts, Inc., 1951), p. 76.

[2]Florence L. Goodenough, "An Early Intelligence Test", Child Development, V (March, 1934), 13-18.

[3]J. McKeen Cattell, "Mental Tests and Measurements", Mind, XV (July, 1890), 375-381.

and others to measure intelligence by means of physical characteristics, sensory acuity, and motor skills tests gave, for the most part, negative results.[1]

During the same period, Binet and his colleagues were experimenting in France with tests of a somewhat similar but less specific type. In 1895, Binet and Henri described ten types of tests which differed from American tests mainly in the much greater complexity of behavior they would measure, and which they thought were likely to discriminate between levels of mental ability.[2]

---

[1]Frank N. Freeman, Mental Tests: Their History, Principles and Applications (Boston: Houghton Mifflin Co., 1939), p. 58.

[2]Anastasi and Foley, Differential Psychology p. 15.

CHAPTER III

THE FIRST FIFTY YEARS OF THE TWENTIETH CENTURY

Thorndike brought out the first book dealing primarily with mental and educational measurements in 1904,[1] and both through this book and his later influence on his students became, more than any other person, responsible for the early development and popularization of standardized educational tests.

Stone, a student of Thorndike, published his arithmetic reasoning test in 1908. This was the first standardized achievement test.[2] Thorndike, in 1909, published his Scale for Handwriting of Children - the first standardized achievement scale.[3] During the period 1909 to 1915, a series of arithmetic tests and five scales for measuring abilities in English comp-

---

[1]Edward L. Thorndike, An Introduction to the Theory Of Mental and Social Measurement (New York: Teacher's College Columbia, 1904).

[2]Cliff W. Stone, Arithmetical Abilities and Some Factors Determining Them (New York: Teacher's College Columbia, 1908).

[3]Edward L. Thorndike, "Handwriting", Teacher's College Record, XI (March, 1910), 83-175.

osition, spelling, drawing and handwriting were published.[1] It is interesting to note that only two of these pioneer instruments were tests, while the remaining five were scales.

Educators at first opposed the standardized test and derided the testers. However, the spread of standardized testing continued, under the stimulation of at least three early developments.

1. The numerous important studies of the accuracy of school marks, revealing the fact that they are highly subjective and inaccurate, demonstrated the need for instruments that would yield more accurate measures of achievement.

2. The surveys of certain of the larger school systems both stimulated the construction and use of tests and were influenced by the development of more objective devices for measuring the abilities of pupils.

3. The development of educational measurements in research bureaus organized in many of the larger school systems, universities, and state departments of public instruction was influential in

---

[1] C. W. Odell, Educational Measurements in High School (New York: Century Co., 1930), pp. 34-35.

popularizing the use of educational tests. Although most of the early standardized tests were for use in the elementary school, it was not many years until the high school and even the college were well provided with such instruments.

The idea of the informal objective examination, referred to during its early days rather loosely as the "New-Type Test" and the "Objective Test", apparently was first publicly expressed by McCall,[1] whose article in 1920 first suggested that teachers do not need to depend solely upon standardized tests, but that they can construct their own objective tests for classroom use. The pioneer book dealing almost entirely with this testing adaptation was published in 1924.[2] The informal objective test has since come into such wide use that a survey in 1936 of testing practices among 1600 high school teachers widely distributed throughout the United States showed that seventy-four per cent used the informal objective and an additional ten per cent used a combination of the informal ob-

---

[1] William A. McCall, "A New Kind of School Examination", Journal of Educational Research, I (January, 1920), 33-46.

[2] G. M. Ruch, The Improvement of Written Examinations (Chicago: Scott, Foresman and Co., 1924).

jective and essay examination.[1]

The history of achievement measurement since the late twenties has been characterized mainly by increasing recognition of the fact that test results offer only one, although the major one, of the types of acceptable evidence on pupil achievement. This tendency toward evaluation, which is broader in scope than testing, has been accompanied by a strong trend toward more scientific use of measurement tools.

Although the contributions of Tyler have been significant in both the standard of testing and the informal objective testing movements, it is probably the latter field that his influence was first felt. He outlined steps of procedure for test construction and validation which clearly pointed out the essential dependence of a program of achievement testing on the objectives of instruction and the recognition of forms of pupil behavior indicating attainment of the desired instructional outcomes.[2] Perhaps he, more

---

[1] J. Murray Lee and David Segel, Testing Practices of High School Teachers (Washington: U.S. Government Printing Office, 1936), p. 6.

[2] Ralph W. Tyler, "A Generalized Technique for Constructing Achievement Tests", Educational Research Bulletin, VIII (April, 1931), 199-208.

than any other test specialist, was responsible for the extension of achievement testing to the more intangible outcomes of instruction, for his contributions thirty years ago doubtless did much to bring into being the modern broad conception of evaluation to replace the earlier and narrower concept of testing.[1]

The eight year study of member schools of the Progressive Education Association, completed in 1942, affected measurement and evaluation practices markedly. The evaluation staff, working under Tyler's direction, developed a series of instruments for measuring such outcomes as logical reasoning, ability to apply principles in the sciences, ability to interpret data, and ability to interpret literature.[2] These and other instruments since made available, including those developed in the Cooperative Study of General Education, are designed to measure functional and relatively intangible outcomes in areas of behavior rather than more formal and intangible instructional outcomes in separate subjects or areas of the curriculum. A related

---

[1] Ralph W. Tyler, Constructing Achievement Tests (Columbus: Ohio State University, 1934).

[2] Eugene R. Smith, et al, Appraising and Recording Student Progress (New York: Harper and Brothers, 1942).

trend is evidenced in the batteries of tests developed during the late 1940's and early 1950's for the measurement of general educational development.

Paralleling the development of paper and pencil tests has been the development of other evaluative tools and of techniques for measuring procedures involved in and products resulting from certain types of skill performances and various other aspects of behavior of the whole child. Prominent among such evaluative tools are the check list, the rating scale, the questionnaire, the pupil profile, the class record sheet and the cumulative record card. Evaluative techniques are represented by the antecdotal report, the interview, the case study, the sociogram and observational analyses of group dynamics.

Binet and Simon brought out the first intelligence scale in 1905, devising it primarily for the purpose of selecting mentally retarded pupils who required special instruction. This pioneer individual intelligence scale utilized the basic idea of interpreting the relative intelligence of different children at any given chronological age by the number of tests of varied types and increasing levels of difficulty they could pass. These characteristics were all re-embodied in the 1908 and 1911 revisions

of the Binet-Simon Scale and also are basic to most individual intelligence scales even today. The 1908 revision introduced the fundamentally important concept of mental age (M.A.) and provided means for obtaining it.[1]

Goddard, Kuhlmann and Terman all adapted the Binet-Simon tests to use with American children during the period from 1911 to 1916. Terman and his collaborators made the Stanford Revision of the Binet Scale available in 1916, and in 1937 followed it with a second and more complete revision. A recent revision, 1960, saw few alterations in content material but a complete restandardization of the conversion tables was carried out. These tests make use of the intelligence quotient (I.Q.), based on the relationship between a child's mental age and his chronological age.[2]

The deviation I.Q. was introduced by Otis in his Otis Self-Administering Tests of Mental Ability which appeared in the late twenties. This concept

---

[1] Freeman, Mental Tests: Their History, Principles and Applications, pp. 86-88.

[2] Ibid, p. 101.

involves a quotient, namely, the ratio of the standard deviation (arbitrarily established) of the brightness measure to the standard deviation of the best scores. This I.Q. concept was subsequently used in the later series developed by Otis, known as Quick-Scoring Tests of Mental Ability. The same technique, but in a more precise form, was adopted for the Pintner General Ability Series and, later, for the revised Terman-McNemar Test of Mental Ability.[1]

The deviation I.Q. had been used on the tests named above for many years before the Wechsler-Bellevue Intelligence Scales, individual tests of intelligence which utilize this concept, appeared. The authors of the Kuhlmann-Finch Test, the Lorge-Thorndike Tests, the revised Kuhlmann-Anderson Tests, and the 1960 revision of the Stanford-Binet Intelligence Scale have more recently adopted the deviation method. It seems clear that the survival of the ratio I.Q. is only a matter of educational lag, and that it will eventually disappear.

---

[1] Walter N. Durost and George A. Prescott, Essentials of Measurement for Teachers (New York: Harcourt, Brace and World, Inc., 1961), p. 75.

Although various psychologists had been working on a group intelligence test, and Otis was near the point of issuing such a test around 1917, the Army Alpha test, used for measuring and placing American army recruits and draftees during World War I, was the first group intelligence test to be published. The Army Alpha Test, widely used for testing men who could read and understand English, was accompanied by Army Beta, a non language test for use with illiterates and men, who, although literate in a foreign language, could not read English.[1] These tests were widely used by educators after the close of the war.

Group intelligence tests began making their appearance almost immediatly following the end of World War I, and the period from 1918 to the middle twenties was marked both by the publication of many such tests and by an upsurge of general interest in intelligence testing. Although the testing techniques have been refined considerably since then, the past quarter century had brought no outstanding changes in the methods of measuring general intelligence. The Army General Classification Test and the Army Individual Test of Mental Ability served functions in World War II closely similar

---

[1] Ibid, pp. 113-135.

to those of Army Alpha and Army Beta in World War I. Several other armed service branches developed comparable instruments for use in their programs of selection and classification.

The measurement of aptitudes, or those potentialities for success in an area of performance that exist prior to direct acquaintance with that area, has been tied up with intelligence testing both fore and aft. Early attempts to measure general intelligence were by means of tests of many specific traits and aptitudes, but that approach was dropped when Binet showed that tests of more complex forms of behavior were superior. It was soon apparent, however, that general intelligence tests were not highly predictive of certain types of performance, especially in the trades and industries.

Munsterberg's aptitude tests for telephone girls and street car motormen in 1913 were followed by tests of mechanical aptitude, musical aptitude, clerical aptitude, and aptitude for various subjects of the high school and college curriculum prior to 1930.[1] Spear-

---

[1] Goodwin Watson, The Specific Techniques of Investigation: Testing Intelligence, Aptitudes and Personality (Bloomington: Public School Publishing Co., 1938), pp. 365-366.

man's splitting of total mental ability into a general factor and many specific factors[1] had its influence on this movement, and accounted for the fact that aptitude tests are frequently called specific intelligence tests.

With the development of factor analysis methods, largely within the last thirty years, certain group factors of intelligence thought to differ from the specific factors or aptitudes and also from general intelligence have emerged.[2] These were first recognized in measurement practice by the introduction of separate linguistic and quantitative, or verbal and non-verbal, scores into certain tests of mental ability that continued to furnish a general measure of intelligence. In addition, several batteries for the measurement of primary mental abilities, differential aptitudes, and general aptitudes, each designed to distinguish several group factors of intelligence, made their appearance during the forties and early fifties.

Impressionistic methods of judging personality and of analyzing character have doubtless been in vogue

---

[1] Charles Spearman, "General Intelligence Objectively Determined and Measured", American Journal of Psychology, XV (April, 1904), 201-293.

[2] Godfrey H. Thomson, The Factorial Analysis of Human Ability (Boston: Houghton Mifflin Co., 1939), p. 14.

for many centuries. They are based in the main on physiognomy, body build or glandular makeup, and divination. Representatives are phrenology, palmistry and graphology.[1]

Personality testing had its antecedents in the work of Kraepelin and Sommer on free association tests during the last decade of the nineteenth century. Although free association tests had persisted to the present day, the questionnaire and rating scale methods used by Galton and others at still earlier dates became the dominant early methods of personality measurement in the United States.[2]

Woodworth devised a Personal Data Sheet, in reality an inventory of neurotic tendencies and emotional maladjustment, for use with American soldiers during World War I. This was probably the outstanding early contribution in this field.[3] A significant number of these structured personality inventories have been developed during the past four decades for the

---

[1] Garrett, Great Experiments in Psychology, pp. 175-181.

[2] Anastasi and Foley, Differential Psychology, p. 22.

[3] Watson, The Specific Techniques . . . . , p. 368.

measurement of adjustment, attitudes, and vocational interests.

Jung, in 1905, published a free association test designed to reveal emotional complexes.[1] Hartshorne, May and their colleagues made exhaustive studies of conduct in largely unstructured or free response situations in the Character Education Inquiry.[2] Although the Rorschach, the modern projective test was introduced in 1921, it was not until 1938 that projective techniques employing such unstructured situations as ink blots and pictures came into wide use in the study of personality.[3]

---

[1]Ibid, p. 368.

[2]Hugh Hartshorne, et al, Studies in the Nature of Character (New York: MacMillan Co., 1928, 1929, 1930), Vol. I, II and III.

[3]Watson, The Specific Techniques . . . . , p. 369.

CHAPTER IV

RECENT TRENDS AND FUTURE DEVELOPMENT OF EDUCATIONAL EVALUATION

It is difficult to gauge the impact of the many developments in educational testing since 1950. Some of these developments are merely advancements of previously established techniques.

One of the most obvious developments has been the increasing use of electronic scoring machines. The many electronic data processing machines now available do not influence testing directly, but they serve as supplements to some of the electronic scoring machines. By recording and summarizing test scores, they facilitate the analysis of results.

Increased financial aid for testing projects such as the National Defence Education Act of 1958, Title V-A, has contributed to the expansion and implementation of testing programs in many schools throughout the United States. In 1955, Nova Scotia, one of the first Canadian provinces to do so, established a provincial standards project. An achievement testing project involving a systematic sampling of pupils in

grades three, six, and nine in successive years. The project has been expanded to include mental ability testing and the results are used as a basis for in-service training for remedial work.

Several publications issued since 1950 have the upgrading of testing as their major purpose. Yearbooks, under the editorship of Oscar Buros, include impartial reviews of standardized tests. These reviews no doubt influence schools in selecting tests, and authors and publishers in planning revisions of old tests and the development of new ones.

Two large scale testing projects deserve special consideration. Project Talent, an American study launched in 1959 has as its aim the expansion of understanding of human talents and the improvement of methods of testing and using test results. The first report of the first stage has been published. The first stage involved the administration of an extensive battery of tests and inventories to some four hundred and forty thousand high school pupils in some thirteen hundred and fifty-three schools representing all fifty states.[1] Many progress reports will undoubtedly be

---

[1] John C. Flanagan, et al, Design for a Study of American Youth (Boston: Houghton Mifflin Co., 1962).

issued between now and 1984, when an extensive follow up of the participants will have been completed. A pilot study reporting results of an international testing project sponsored by UNESCO seems likely to pave the way for future developments in cross-cultural and culture-fair testing.[1]

There has been, and it appears as if there will continue to be, an increase in College Admissions Testing. For years, the College Entrance Examination Board (CEEB) had this field to itself. Recently, the American College Test (ACT) has entered the field, along with National Merit Scholarships Tests. With computers, much more information is being made available to universities, schools and students. The data is processed and school norms, state norms and university norms calculated. Predictions are made for each student as to his expected average at several universities - whether he is a "C" student or better.

Further, the chances of the student earning a passing grade in each of several electives subjects are also calculated and made available to the student and

---

[1] UNESCO Institute for Education, Educational Achievements of Thirteen-Year Olds in Twelve Countries, by Arthur W. Foshay, et al, (Hamburg, Germany: International Studies in Education, 1962).

the counsellor.

We may look forward to an extension of College admissions testing to Canada. McGill and Bishop's College now require the CEEB. The Canadian Universities Foundation has received the report of a committee set up to investigate admissions testing which recommendéd that negotiations be carried out with the CEEB in this matter.

If external testing increases, there may develop a tendency for the schools to discontinue their own testing programs due to duplication. There are those who contend that this is an abnegation of responsibility, since it places emphasis on the college group, while the school must attempt to serve all its pupils.

Several new concepts and emerging trends should be noted along with the specific indications of progress just discussed. Probably the most significant ones have occurred in intelligence and achievement measurement. A recent trend in ability testing is toward distinguishing creativity - variously described as ingenuity, originality, and divergent thinking - from general intelligence. A number of tests (many of which are not yet standardized) have been devised to measure this newly emerging dimension of mental ability. Results of intelligence testing have increasingly been interpreted in terms of a deviation form of intelligence quotient, actually a

standard score with a mean of 100 and a standard deviation of 16, as a replacement for the obsolescent quotient form of I.Q. Two other recent trends are to present test results with accompanying devices for interpreting the reliability of test scores, and to outline shortcut procedures for estimating, rather than computing, the basic statistical measures needed to interpret test results.

Also influencing educational achievement are the recently developed, and still developing, reorganizations of the curriculum - especially in mathematics, physics, and foreign languages. Since the new courses are still in the formative stage and have not replaced more traditional course patterns in many schools, standardized tests do not yet significantly reflect the changes. Achievement tests, however, appear to be increasingly adapted to the measurement of complex instructional outcomes, such as understandings and applications, and to be founded on more careful and extensive analyses of educational objectives and related outcomes.

In the future, increased attention may be paid to differential predictibility of tests for different people. Research findings indicate that a test may have different degrees of validity for different people,

in predicting the same criterion. For example, two investigators reported that the engineer scale on the Strong Vocational Interest Blank had greater predictive power for students who had low accountant score on the Strong than for those who had a nigh score on the accountant scale. Another found that a tapping and dotting test was a much more valid predictor of taxi driving success for those drivers who had a low occupational level than for those with a high occupational level. Thus we may find that certain tests predict well for plodding students but not for others; or for students of lower socio-economic status but not for others. Applications of this, while speculative, are nontheless interesting. At present, the criterion for university or college success is a pass or fail. But it is not inconceivable that it takes a different set of abilities to do well in the second and the third years.

Some years ago, many test users drew a rather clear line between intelligence tests and achievement tests. Today, scholastic aptitude tests and tests of general educational development have become almost indistinguishable. The two seem to be about equally effective in fulfilling one of their major functions; prediction of scholastic success. Both depend mater-

ially upon the results of learning, or on a high degree of educational loading. Toward the other end of a continuum are culture-fair tests that place minimal dependence on learned behavior by using nonverbal materials.

In personality measurement, the present tendency is to explore increasingly the value of devices lying between the highly structured self-report inventories and the relatively unstructured projective techniques. Such techniques in many instances apparently still suffer from a dearth of evidence to support their reliability and validity. Consequently, various situational tests, techniques of interaction, process analysis, and sociometric procedures have been receiving, and seem likely to continue to receive, considerable attention.

CHAPTER V

TESTING: AN EDUCATIONAL INNOVATOR

Standardized tests have served as an innovator for most of the changes that have taken place in North American elementary and secondary education over the past forty years. If the testing movement is to receive substantial credit for these changes from those who approve of them, it must also be prepared for the criticism of those who disapprove.

In the three decades following World War I, the development, ready availability, and widespread use of standardized tests of mental ability and educational achievement provided North American educators with a vast amount of reasonably accurate data about individual differences. These data served to spotlight the need for modifying instructional practices in order to accommodate the wide range of potentiality and performance found in a typical classroom. During the same period, research workers were using standardized tests to accumulate the facts that are the basis for most of our generalizations about mental development, the learning process and the factors

that influence it, group behavior, and the effectiveness of various instructional procedures.

By the late 1930's, it was clearly documented that the mental ages of first graders had a range of more than four years and that the spread increased in higher grades. For example, sixth grade readers could be expected to have mental ages ranging from eight to sixteen; differences as much as ten years in mental age could be found among high-school students in the same grade. Results from standardized reading tests provided evidence that some pupils in the fifth grade could read as well as the typical tenth grader, while some high school seniors scored at the level corresponding to the average for sixth grade pupils. Similar variability characterized achievement in all school subjects. Tests of English, arithmetic, science and history administered to pupils in the eighth grade yielded achievement ages ranging from six to eighteen.

The amassing of facts such as these - all pointing to important differences among pupils - inevitably led to a re-examination of educational principles and practices. Homogeneous grouping, differentiated goals, promotion and retardation, variegated course offerings, individualized and remedial instruction - all came to be seen as centrally relevant issues in

light of the knowledge derived from standardized tests. Educational philosophers, curriculum planners, administrators, and specialists in teaching methodology, quickly turned their attention to analyzing tne consequences of these facts and to suggesting courses of action.

Test authors were aware of the educational implications of the data provided by their instruments. The manuals for standardized tests and the textbooks for basic courses in educational measurement increasingly reflected the view that the need to eliminate uniform treatment of students and lock step teaching methods was the primary inference to be drawn from the variance innerent in test scores. Increasing emphasis was placed on using test data, and not merely recording them. Suggested uses frequently necessitated changes in school practices. Thus, the manuals for achievement tests advised administrators and teachers to use test results for placing students in groups, for counselling students about educational and vocational plans, for diagnosing individual learning difficulties, for evaluating continuously the appropriateness of objectives, for motivating learning, and for appraising and modifying instructional methods and emphases. In effect, a relatively complete operation-

al philosophy of education and a firm set of mutually compatible values were inherent in the practices recommended.

A test maker's educational creed can be deduced quite readily from the kinds of school purposes and uses that he suggests for his instrument. Such an examination shows that the developers of standardized achievement and aptitude tests regard the ideal school as one in which the curriculum has breadth and flexibility - so that there can be realistic tailoring of programs to student abilities, needs and aspirations; one in which teachers and administrators continually gather and analyze data about their students - so that objectives and the variety of learning experiences offered will be fitting and adequate; and one in which there is strong commitment to the child-development point of view - so that pupil performance may be interpreted within a framework of established cause and effect relationships. It was the espousal of such educational ideas by test workers, coupled with the general failure of schools to use test data for many of the recommended purposes, that prompted Cook, as early as 1951, to summarize the situation thus: "educational measurement has outrun educational practice".[1]

---

[1]Walter W. Cook, "The Functions of Measurement in the Facilitation of Learning", Educational Measurement, IV (April, 1951), p. 45.

In the intervening years, many schools have worked to close the gap by putting into practice more and more of the procedures that are necessary if maximum use is to be made of standardized test results. Such procedures most always involve more effective provisions for meeting individual differences. They frequently necessitate reduction in class size, broadening of course offerings, an increase in counselling personnel and facilities, the employment of specialists in remedial instruction, and the purchase of a wide variety of instructional material and aids. Critics of modern education, many of whom believe that schools were best when they ignored individual differences, and economy-minded citizens who wail about the expenses of modern schools, stand united in their opposition to such innovations.

The testing movement has been one of the primary forces contributing to the development of the best practices of contemporary education. By insistently and repeatedly marshalling the array of individual differences, it has knawed at the nation's educational conscience. Because of its influence on contemporary education, testing cannot escape criticism and condemnation from those who regard our schools as ineffective, or as being undermined by costly and unnecessary frills.

It is perhaps not without significance that the Bulletin of the Council for Basic Education carried a highly critical and misleading review of Lindquist's Educational Measurement seven years after the book's publication.[1] The critics may be late in discovering the impact of the testing movement on educational practice, but they are not likely to overlook it.

Although the influence of the testing movement on educational procedures has been widespread, there are still many schools where the impact has been minimal. Far too many teachers and administrators are still vexed by the question: "What really practical use is to be made of the results from a standardized testing program"? The utilization of test data in a school is intimately related to the school's philosophy and to the amount of research orientated thinking that obtains in the school. Standardized tests have been developed by men who do not consider education to be a closed topic - one for which all the final data are now available; the effective use of their instruments can be achieved only in settings where there is at least a basic sympathy with this view. In school situations

---

[1] Edmund Gibson, "A Review", Educational Measurement, II (June, 1958), p. 9.

where student failings are seen as reflections on the student only - situations where there is no disposition to question goals or to re-examine them in face of changing student characteristics and ambitions, where instructional procedures and curriculum are time honored - standardized test data would be of only very limited value. On the other hand, teachers and administrators who are conscious of the need for improvement and who see educational practices as undergoing a process of evolutionary refinement, and not as procedures that are immutably fixed, find uses for test data that provide both enlightenment and the spur to action.

Recommendations from testing specialists concerning the ways in which schools may utilize test results are based on a conception of school procedures that is in advance of the actual situation in many schools. Consequently, existing measures of academic ability, achievement, aptitude, and interest place in the hands of school personnel vastly greater quantities of information about students and the effectiveness of the educational program than they are able to use to maximum advantage because of rigidly limited student options and teacher practices. Our standardized tests are twentieth-century tools for twentieth-century schools; the updating of school procedures is the necessary condition for im-

proved utilization of test data. Fortunately, even the limited and fragmentary use of tests and test data contributed to the attainment of that condition.

The recent expansion of guidance and counselling services in North American high schools is due to a large measure to the advances made in the area of standardized tests. As tests became more numerous, and the use and interpretation of test data - including the development of local norms, expectancy tables, and regression equations - became more sophisticated, school administrators clearly recognized that specialized skills were needed if full advantage was to be derived from the standardized testing program. Thousands of new counsellor positions were created in the American high schools, largely because of the pressures created by the testing movement. The work of guidance counsellor, in turn, has given tremendously increased visibility to tests.

It is a strange paradox that our elementary schools, which engage in a much greater volume of internal testing using standardized materials than our high schools, have not felt a similar need to employ staff members with specialized competence in testing. Instead classroom teachers bear the responsibility for obtaining the information about pupil performance that is provided

for the student and his parents. As parents of elementary school children come to realize the frequency with which their children are given standardized tests, and as they come to understand more fully the effects that test performance in the elementary grades may have on the future educational careers of their children, they will undoubtedly exert pressure for improved test interpretation, for more insightful analysis of pupil strengths and deficiencies, and for more attention to individual needs. Once again, testing will have served as an innovator for modifying educational practice.

The visibility of internal testing programs was severely limited as long as schools took the position that student scores were privileged information and that the communication of test results to pupils and parents would result in more harm than good. In recent years, there has been a marked trend toward informing the individual and his parents about test performance. Increasingly, schools are guided by the proposition that parents have a right to know what the school knows about their children, but that the school has an obligation to communicate the information in a form that will minimize the possibility of erroneous interpretation.

In the United States, millions of American

families have now had first hand knowledge of such documents as the pupil and parent report forms for the Iowa Tests of Educational Development, the percentile band profiles of the SCAT and STEP series, and the bar graph profiles of the Differential Aptitude Tests. Terms, such as percentile, standard score, error of measurement, normal distribution, validity, reliability, norms, stanines, and equivalent forms are becoming familiar to parents as well as students. Even popular magazines are attempting to assist people in their understanding of the nature, limitations, and uses of standardized tests.

Parents and students clearly recognize that test results have consequences; the increasing visibility of tests in our schools and in other aspects of our society has served to heighten the concern about tests and the uses made of test data. The effects of standardized tests on the pupil and his family are likely to be direct and far-reaching, but empirical research on the social consequences of testing has been lacking. Goslin's recent book details many of the questions relating to the impact of testing for which we do not have adequate answers.[1]

---

[1]David A. Goslin, The Search for Ability (New York: Russell Sage Foundation, 1963).

Nonetheless, the growing visibility of standardized tests as an important part of school practice has alerted parents and pupils to the possibility that an individual's subsequent course of development may be vitally affected by his performance on these instruments. Outstandingly high or disappointingly low scores may have an impact on the student's self-concept; on his relationship with parents, siblings, peers, and teachers; on his adjustment; and on the educational and vocational opportunities open to him. His own decisions and those that others make about him are likely to be influenced by his test performance. Goslin believes:

> Repeated experiences with objective tests may cause an individual to fear or resent the invasion of his privacy that a test may represent, particularly in those cases where the test is perceived as a barrier to reaching of a goal rather than a means for its attainment. Anxiety, in turn, may influence succeeding test performances, and thus, in effect, reinforce itself. It has been pointed out that as long as there is debate over the validity of objective tests, people will have a means of rationalizing a poor performance. On the other hand, as tests are improved technically, and are accepted as valid indices on an individual's abilities (and are therefore depended upon more in making decisions), anxiety over testing may reach a much higher level.
>
> . . . . . .
>
> As people become aware of the uses that are being made of tests, it is possible that the constitutionality of intelligence testing will be questioned, particularly in those cases where test scores are misused. In any event, it is likely that questions such as these will be argued more and more frequently as our educational system moves in the direction of greater specialization and over-all excellence.[1]

---

[1] Ibid, pp. 183, 185.

While reporting the results of internal testing programs has added visibility to school testing programs and, therefore, concern - a healthy concern - about the consequences of test scores, such programs have not ordinarily produced controversy. Instead, most of the heated discussions of testing in recent years have stemmed from external testing programs, such as the College Entrance Examination Board tests, the American College Testing Program, or the National Merit Scholarship tests.

External tests have proliferated since 1950. Tyler attributes their increase in scope and magnitude to four factors:

1. The increase in the number of students wanting higher education, and the resultant increase in competition for the available spaces.

2. The larger number of students applying for admission to distant institutions, and the attendent difficulty of evaluating transcripts from high schools that are not well known to the admissions officer.

3. The considerable increase in scholarship funds granted on the basis of competitive examinations.

4. The development of high speed electronic scoring equipment which has made possible the rapid, accurate, and economical scoring of large numbers of answer sheets for such centralized testing programs.

Tyler notes that:

> Although each of these developments is a good thing, many schools were unprepared for them, and the resulting effects upon the schools have, in many cases, created problems.[1]

External testing programs provide the basis for immediate and direct decisions to admit or not to admit, or to give or withhold scholarship assistance to a particular candidate - even when test data are used in conjunction with other types of information about the applicant. The impact of such programs, not only on students but also on their teachers, counsellors, principals and superintendents, is strong and pervasive; schools - as well as individuals - compete for the recognition that comes as a result of high scoring performance. External testing programs thus have a built-in quality of high visibility. It is understandable that such programs, which involve high stakes and manifestly affect the lives of many people, should spawn anxiety, unrest and criticism in at least some of those who are directly affected by the results.

These external testing programs, which so many students, parents and educators now regard as critical and all-important experiences, have been

---

[1] Ralph W. Tyler, The Impact and Improvement of School Testing Programs (Chicago: The Society Press, 1963), pp. 194-195.

challenged frequently on two points. It has been contended, in the first place, that the multiplication of such testing programs is unnecessary and that this over-testing results in wasted time for students and members of the school staff. At best, this is a peripheral criticism and one that fails to consider the many possible advantages of variegated external testing programs. Risking all of a student's opportunities for admission or scholarship aid on a single test may have virtue from the point of view of administrative efficiency, but such efficiency can hardly be the summum bonum. Research and the development of more refined measurement techniques are likely to increase the differences, rather than the similarities, in tests used by various external agencies for their distinct purposes. Beyond that, competition in test development - as in most areas of human activity - tends to ensure greater progress than is achieved in a monopoly arrangement. From the candidate's point of view, gaining familiarity with a wide variety of external tests is a beneficial educational experience and one that takes less time than he might spend in glee-club rehearsals or football practice.

The second type of criticism directed against external testing programs is the charge that scores derived from these tests are often misused. There is ev-

idence to support this charge. It includes the unethical and invalid practice in some school systems of releasing test results to local news media in order to support the claim of superiority over a neighboring school system. It also included the establishment and arbitrary application of critical cutting scores (ignoring standard errors of measurement) for college or graduate school admission or for the awarding of financial assistance. But this kind of legitimate criticism is directed not against the tests themselves, but against the improper use of the data derived from these instruments.

In their future development and use, external testing programs are likely to produce changes in many aspects of higher education, as internal testing programs have already promoted innovations in curriculum and instruction in elementary and secondary schools. If aspects of creativity are to be included in college admission tests, or if efforts are to be made to match institutional characteristics with student needs, external testing programs may well prepare the ground for revolutionary changes in the structure and character of higher education.

Test data have consequences. While educational agencies may not at first be willing to face up to these

consequences, the continuing accumulation of data increases the probability that their implications will not always be ignored. The impact of the testing movement is widespread; it has both personal and societal repercussions. Standardized testing has become a prominent and active agent for the promotion of change in our evolving culture. Critics and defenders of the testing movement may yet find that their battle line is drawn between and camps of those who oppose and those who favor using every reasonable means to improve our way of life, including the schools.

BIBLIOGRAPHY

Books

Anastasi, Anne and Foley, John P. Differential Psychology. New York: MacMillan Co., 1949.

Ayres, Leonard P. History and Present Status of Educational Measurement. Bloomington: Public School Publishing Co., 1918.

Boyd, William. The History of Western Education. London: A. and C. Black Ltd., 1921.

Caldwell, Otis W. and Courtis, Stuart A. Then and Now in Education 1845-1923. New York: World Book Co., 1923.

Durost, Walter N. and Prescott, George A. Essentials of Measurement for Teachers. New York: Harcourt, Brace and World Inc., 1961.

Flanagan, John C., et al. Design for a Study of American Youth. Boston: Houghton Mifflin Co., 1962.

Freeman, Frank N. Mental Tests: Their History, Principles and Applications. Boston: Houghton Mifflin Co., 1939.

Garrett, Henry E. Great Experiments in Psychology. New York: Appleton-Century-Crofts Inc., 1951.

Goslin, David A. The Search for Ability. New York: Russell Sage Foundation, 1963.

Hartshorne, Hugh, et al. Studies in the Nature of Character. Volume I, II and III. New York: MacMillan Co., 1928, 1929 and 1930.

Knight, Edgar W. Twenty Centuries of Education. Boston: Ginn and Co., 1940.

Lang, Albert R. Modern Methods in Written Examinations. Boston: Houghton Mifflin Co., 1930.

Lee, Murray J., and Segel, David. Testing Practices of High School Teachers. Washington: United States Government Printing Office, 1936.

Martin, W. A. P. The Chinese: Their Education, Philosophy, and Letters. New York: Harper and Brothers, 1881.

Odell, C. W. Educational Measurements in High School. New York: Century Co., 1930.

Peterson, Joseph. Early Conceptions and Tests of Intelligence. New York: World Book Co., 1925.

Smith, Eugene R., et al. Appraising and Recording Student Progress. New York: Harper and Brothers, 1942.

Stone, Cliff W. Arithmetical Abilities and Some Factors Determining Them. New York: Teacher's College Columbia, 1908.

Thomson, Godfrey H. The Factorial Analysis of Human Ability. Boston: Houghton Mifflin Co., 1939.

Thorndike, Edward L. An Introduction to the Theory of Mental and Social Measurement. New York: Teacher's College Columbia, 1904.

Tyler, Ralph W. Constructing Achievement Tests. Columbus: Ohio State University, 1934.

_____ . The Impact and Improvement of School Testing Programs. Chicago: The Society Press, 1963.

Ruch, G. M. The Improvement of Written Examinations. Chicago: Scott, Foresman and Co., 1924.

Russell, Charles. Standard Tests. Boston: Ginn and Co., 1930.

Watson, Goodwin. The Specific Techniques of Investigation: Testing Intelligence, Aptitudes, and Personality. Bloomington: Public School Publishing Co., 1938.

Articles and Periodicals

Cattell, J. McKeen. "Mental Tests and Measurements", Mind, XV (July, 1890), 375-381.

Chadwick, E. B. "Statistics of Educational Results", The Museum, a quarterly magazine of education, literature and science, (January, 1864), iii, 480-484.

Cook, Walter W. "The Functions of Measurement in the Facilitation of Learning", Educational Measurement, IV (April, 1951), 45.

Gibson, Edmund. "A Review", Educational Measurement, II (June, 1958), 9.

Goodenough, Florence L. "An Early Intelligence Test", Child Development, V (March, 1934), 13-18.

McCall, William A. "A New Kind of School Examination", Journal Educational Research, I (January, 1920), 33-46.

Spearman, Charles. "General Intelligence Objectively Determined and Measured", American Journal Psychology, XV (April, 1904), 201-293.

Thorndike, Edward L. "Handwriting", Teacher's College Record, XI (March, 1910), 83-175.

Tyler, Ralph W. "A Generalized Technique for Constructing Achievement Tests", Educational Research Bulletin, VIII (April, 1931), 199-208.

Other Sources

Old Testament. Holy Bible. Judges 12: 5-7, Cleveland: The World Publishing Co., 1947.

UNESCO Institute for Education. Educational Achievements of Thirteen-Year olds in Twelve Countries, by Arthur W. Foshay, et al. Hamburg: Germany International Studies in Education, 1962.