EDUCATIONAL TESTING AND MEASUREMENT: A BRIEF HISTORY

David L. McArthur

CSE Report No. 216
1983

Center For The Study of Evaluation
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Educational assessment in the Western tradition has a long but very irregular history. Seven centuries ago, one English college was deemed remiss in its responsibilities because its founder had determined that its recent graduates "...expressed themselves very inaccurately in the learned languages..." (Sylvester, 1970, p.19) the method of such determination was not described. A tradition of oral examinations was built up over several centuries, only to disintegrate almost completely by the time Isaac Newton attended college about 1660; not only were there no examinations but frequently the lecturers themselves simply never showed up for classes. However, in another hundred years, both Oxford and Cambridge, recognizing the deteriorated situation, decided to improve their curriculum and instituted regular written examinations in a variety of topics. The exams of this era were almost exclusively essay questions emphasizing factual recall; one extant example shows eight questions each in history and geography, and six in grammar, primarily Latin and Greek. In the education of the younger pupils, examinations began to become more prevalent as textbooks for the grammar school came to be formulated into distinct grade levels.

> The new sequences of textbooks allowed a more precise grading to be implemented in schools in various parts of Europe...Within the school a further step was the development and application of the principle of a child's regular progression through grades at various intervals of about a year (Bower, 1975, p.419).

The Jesuits, finding that such a procedure fit perfectly into their concept of ratio (the systematically ordered body of knowledge) took up the idea with vigor, and it rapidly spread across Europe.

Meanwhile, in China, civil service examinations were already several millenia old. The earliest proficiency testing dates from 2200 B.C., and formal procedures for examination date from 1115 B.C. Despite a concentration on literary rather than managerial skills, the system was to be the model for a number of efforts at standardizing competition for civil service positions in Europe and the U.S. during the 19th century. But in China the testing system was abolished in reforms at the beginning of the 20th century, as Western technologies and educational orientations intruded into the Orient (DuBois, 1964, 1967).

In the United States, it was not until 1845, following Horace Mann's advocacy of written examinations, that testing was incorporated into educational practice. The first recorded examination was administered in Boston that year, and the concept took hold quickly (Englehart, 1950). Within thirty-five years, promotion from grade to grade was no longer made by personal recommendation but instead invariably was judged by success or failure, scored as a percentage, on a written exam. Mann's viewpoint of testing, while not using the word "objective," carried with it a decided bias towards objective measurement and standard tests (Ruch, 1929). The earliest objective educational tests are found in a book complete with questions, answers and scales, by an English schoolmaster, dated 1864 (Kelley, 1927). Objective tests in spelling and arithmetic were in place in the U.S. by the 1870's. Then, in 1881, the superintendent of schools in Chicago, expressing a strong sentiment against testing in particular

(if not against science in general) decreed that advancement of students was to be carried out only by direct recommendations of teachers and principals. Testing for purposes of grade-level advancement was prohibited. His viewpoint was widely shared; suddenly, the impetus for "objective" measurement and assessment was on the wane. "Examinations for grade promotions were gradually abolished in all the best schools," claimed the superintendent's successor. "The person best qualified to judge of a child's ability to go on is his teacher...To say that any other test is necessary is a travesty on common sense" (Bright, 1895, pp.274-275). By the end of the nineteenth century, educational testing had achieved a bad name. Teachers were "teaching on the test," devoting weeks of preparation and drill to extant editions of upcoming exams, and the public was not pleased.

A completely separate thread in the fabric of educational measurement is found in a review of the history of statistics. The first lectures in statistics date around 1660; the first use of the word "statistic" is placed at 1749, in reference to the accounting of all the things that make up a kingdom (Meitzen, 1891). While extensive developments in mathematics were being made during this time (Newton, for example, was solving problems in differential calculus by 1676), the setting out of facts and figures in the social sciences for many years was limited to tabulations of various facts, actuarial tables, and census taking, the first about 1769 in Denmark. Interestingly, some recognition of the importance of understanding

individual differences in mental abilities is found in the field of astronomy by 1822 (Freeman, 1926). It was not until this century that the word "statistics" came to refer exclusively to quantitative approaches; its origins apparently are tied to the Germanic discipline called "Staatenkunde" or study of governments and politics. The profession suffered a decline as the old teachers passed away, and the task of statistics was made increasingly narrow.

> In 1806 and 1807 a passionate controversy arose against the brainless bungling of the number statisticians, the slaves of the tables, the skeleton-makers of statistics...The opponents in the sharp attack were themselves, however, not sufficiently clear how new and precise limits for their science should be determined. (Meitzen, 1891, pp.49-50).

An International Statistical Congress was formed to attempt to resolve the confusion; it met first in 1853 and showed a surprising degree of success. Even though its members chose to stay out of issues of statistical theory, in 1869 one of their resolutions declared:

> ...that in all statistical researches it is important to know the number of observations...; the qualitative value is to be measured by the divergences of the numbers among themselves as well as the average...; it is desirable to calculate...the average deviations (Meitzen, 1891, p.80).

These principles formed the basis for technical developments in educational statistics into the twentieth century: one of the first texts (Rugg, 1917) devoted most of its efforts to tabulation, averages, frequencies and variabilities. Despite several pioneering studies in educational attainment, in large measure the collection and analysis of data at this time was confined to tabulations of school attendance and costs. The statistical societies of the day were deeply embroiled in social problems, especially the relations of education to

crime, and spent no time at all on assessing educational achievement beyond such indices as the ability to sign one's own name (Cullen, 1975).

By the middle of the nineteenth century, considerable progress had been made in the analysis of experimental data from agricultural research. Good experimental designs, including factorial and split-plot techniques, were in place about 1850. Galton spent time investigating how mathematical solutions might best be developed for data from studies of Charles Darwin, building a number of statistical tools in the process, and was the first to attempt measuring characteristics of individual intelligence (1883). But it was not until Pearson's chi-square test (1900), and Student's t-test (1908) that appropriate quantification of educational data could be developed, although the latter, surprisingly, took a number of years to catch on (Cochran, 1976). Fisher's analysis of variance (1924) drew heavily on these precursors but it too was relatively slow in being incorporated into the repertoire of educational statisticians. Guilford's text on fundamental statistics in 1942 awards analysis of variance fewer than nine pages, embedded in a chapter on reliability.

In 1890 appeared the first study of reliability (Edgeworth, 1890). In the same year the seminal short article by Cattell (1890) marked the first time the words "mental tests" were used together. Following Galton's lead, several investigators in Germany began to develop mental tests, and in the U.S. there was extensive interest in the relationship of mental capacities to physical characteristics.

The American Psychological Association set up a standing committee in 1895 to consider cooperative efforts in mental and physical statistics; the American Association for the Advancement of Science did likewise the following year. Binet, who had been working on problems in mental reasoning since 1886, wrote an important article in 1898 on the utility of measurement and scaling in the appraisal of human intelligence. However, two major studies of testing around this time (Sharp, 1899; Wissler, 1901) concluded that many of the available tests used for psychological research fell far short of their claims, in both content and method (Peterson, 1925). In education, Rice's (1897) study of spelling attainment, using a single list of 50 words in a test administered to 30,000 children, was a pioneering study, which circulated widely but gained few supporters (Wilds & Lottich, 1970).

About the turn of the century there was a fair degree of public discouragement about educational testing. However, about this time, the first survey of school facilities and educational practice was conducted, the College Entrance Examination Board was established, and in 1902 the first course in educational measurement was taught (by Thorndike at Columbia) (Meyer, 1965). Concurrently, interest in the concept of general intelligence was being pursued by a number of investigators, following a suggestion by Galton in 1883 and a study of 1,500 children conducted in 1891 (Burt, 1909). In the analysis of results from the latter investigation, however, came the explicit realization that statistical methods for educational measurement were

in desperate need of thoughtful improvement. Burt speculated that the consistent failures of research investigations in the area of general intelligence before the turn of the century

> were largely due to their reliance for discovery of correlations upon mere inspections of the data they obtained, instead of upon quantitative determination and mathematical deduction (pp.94-95).

During the first decade of the twentieth century, the growing impetus for increased statistical rigor could be felt in several areas; measurement successes in anthropometry and biology provided much needed support for such improvement. In 1904, Toulouse and Pieron's two volume manual on laboratory experiments included sections on intelligence and the measurement of individual differences. In 1906 the American Psychological Association created a permanent committee charged with evaluating requirements for standard laboratory technique and appraising both group and individual tests with attention to practical applications. Binet's test for intelligence (1905) and Thorndike's book on mental measurement (1904) had particular significance during this time, as did Spearman's (1904) paper on general intelligence. By 1910, a vast number of tests in skills like English, spelling, handwriting, reading and arithmetic had emerged, followed closely by more technical articles on topics like numerical analysis, standardization, validity and correlations.

> ...American educators quickly realized that the scale idea could be applied not only to intelligence but to achievement as well. There followed a phenomenally creative period during which testmakers developed instruments for virtually every aspect of educational practice (Cremin, 1961, p. 186).

In 1913, the National Council of Education released a major report on standards and tests for measuring school efficiency, and expressed this sentiment:

> We are only begining to have measurement undertaken in terms of standards or units which are, or may become, commonly recognized. Such standards will undoubtedly be developed by means of applying scientifically derived scales of measurement to many systems of schools. From such measurements it will be possible to describe accurately the accomplishment of children and to derive a series of standards...(Strayer, 1913, p.4).

Graves, reviewing the condition of education in 1913, expressed the sentiment that the application of mathematics to measurements in education was one of the most significant movements of that time.

Developments in objective measurement of intelligence and educational achievement came to a head with the crisis of the Great War. Work in Germany on the screening of inductees had been in progress since 1905; Binet and Simon (1910) discussed the application of intelligence testing in the French army (Peterson, 1925). In the U.S., Terman's revision of the Binet scale was completed by 1917, and was applied soon thereafter to the testing of 1.7 million recruits. A small team of educational psychologists produced the Army Alpha and Beta tests of intelligence between May 28 and June 10, 1917; a copy of the examiner's manual was enroute to the printer within a month. Immediately after the war, as the Army was selling thousands of unused test blanks, both educational specialists and the public began to realize that objective test results had to be taken with some degree of caution. One of the originators of the Army Alpha expressed the

sentiment unambiguously: "We do not know what intelligence is and it is doubtful if we will ever know what knowledge is" (Goddard, 1922, quoted in Spring, 1972, p.5). Even so, by 1920, objective testing formed the core of educational assessment methods. The Journal of Educational Measurement devoted several issues in 1921 to a symposium on scientific measurement of intelligence.

During the decade that followed, the objective assessment of intelligence "swept America, and to a lesser extent Canada, like an educational crusade...The critics were numerous but few in comparison to the advocates..."(Marks, 1976, p.10). McCall's (1922) book on educational measurement and Monroe's (1923) the following year were the first to set out the procedures for a "new type examination," the multiple-choice and true-false tests. Principles of test construction began to earn chapters of their own, and the variety of interpretations and uses of tests was becoming a major consideration for many educators (Monroe, 1945). Then came the first contributions to what is now recognized as classical test theory: Thurstone's (1925, 1926, 1927) articles on the scoring of individual performance, Ruch and DeGraff's (1926) study of corrections for guessing, Ruch's (1929) The Objective or New Type Examination, and Thurstone's The Reliability and Validity of Tests, 1931.

The concept of reliability is illustrative of the historical development of educational measurement. Because of its basis in correlational method, which was already well advanced at the turn of the century, a number of technical articles appeared quite early

concerning the statistical nature of reliability indices. By the time that a major study was launched in the late 1920's by the American Historical Association's Commission on the Social Studies into the nature of testing in social sciences education, reliability measures were regarded as essential by technical specialists but generally disregarded by practitioners. Under the counsel of Truman Kelley, a large-scale investigation was conducted on the use of tests for determining overall class and school performance, recognizing individual skill levels and individual differences, and appraising attitudes and personality traits. It also studied the utility of the "new-type" tests. In the long run both the social science specialists and the educational measurement technicians were disappointed in the results of the study. The former were not pleased by the tendency of short-answer and multiple-choice tests towards fragmentary presentation of, and limitations to, simple facts in the curriculum and the deletion of shades of meaning. The latter felt that lack of objective terms, which they saw as essential for objective measurement, obviated the study's conclusions. Kelley's feelings were sufficiently strong that he wrote a 15-page appendix entitled "A Divergent Opinion as to the Function of Tests and Testing" in which he excoriated the opponents of testing with more than a dozen carefully reasoned arguments regarding the appropriate scientific use of educational tests, plus one or two direct strikes to the more emotional nature of the argument:

> The opponents (of testing) show no awareness of the tests of reliability and validity of measuring instruments, either

> judgments of teachers or of test scores. We believe that such awareness is essential to any educator who is not content to work in the dark (p. 489).

In the areas of reliability and validity, technical proofs were available as early as 1910 (Spearman, 1910) providing a rationale behind error measurement and (Brown, 1910) giving a definition of true score. But it was some time before either term was given serious treatment in the standard texts. Taking a representative contribution from each decade, we find a half-dozen index entries in Rugg's 1917 text, 18 entries between the two in Ruch's 1929 text, four chapters in his 1942 book, and eight full chapters devoted to the two topics in Gulliksen's 1950 text. However, by the 1930's there had accumulated a variety of estimation procedures and a great deal of confusion of terms (Adams, 1936; Barthelmess, 1931; Lincoln, 1932). An attempt to resolve the issues was made in Thurstone's small book on the topic in 1931, another in Kuder and Richardson's (1937) key article on test reliability, followed by Guttman's (1945) reformulation and Cronbach's (1947) discussion of the several different kinds of reliability coefficients. The American Psychological Association tried to resolve the various discrepancies by committee in 1954. Tryon (1957) provided an extensive historical review of the reliability concept and a domain-sampling reformulation. "The extraordinarily massive literature in this topic," wrote Cattell (1964), "...has never lacked statistical finesse and mathematical virtuosity" (p.1), but he, too, felt a need to suggest substantial redefinitions for both reliability and validity, which in turn were ignored four years later with publication of a definitive mathematical analysis by Lord and Novick (1968).

The first formulations of a 'sample-free' approach to mental measurement are found in Lawley's (1943) analysis of item selection. Although the problem had been explored tangentially by Horst (1936) and more recently by Ferguson (1942), his paper was among the earliest to seek mathematically rigorous justifications for the selection of maximally discriminating test items, and to examine in some detail the concept of item characteristic curves. Tucker (1946) provided further statistical support. Gulliksen (1950) summarized the early work in true score theory, and Lord explored the application of latent trait theory to test theory with his doctoral dissertation, published as Theory of Test Scores (1952). Interestingly, he felt that the actual utility of large portions of the theory would be limited in practice by the difficulty in obtaining sufficiently large data sets, and did not publish about the problem again for another ten years. At that point he presented an important development, the beta-binomial model of the frequency distribution of true scores and raw scores (Keats & Lord, 1962), and further refined the definition of true scores in Lord & Novick (1968). Meanwhile, Birnbaum explored certain statistical properties of normal and logistic characteristic functions in 1957 and 1958, but few other papers on this topic appeared until the 1960's.

The sentiment has been expressed more than once that the science of educational testing has progressed fitfully. Despite a plethora of statistical developments, "most of the major theoretical and technical distinctions and most of the principle points of dispute were in existence by 1925" (Thomson & Sharp, 1983). This includes such

diverse topics as item analysis, test bias, the nature vs. nurture arguments regarding individual intelligence, and at least the beginnings of factor structure explanations for educational assessment.

REFERENCES

Adams, H. F. Validity, reliability and objectivity. In W. R. Miles (Ed.), Psychological studies of human variability, Psychological Monographs, 1936, 57, 329-350.

Barthelmess, H. M. The validity of intelligence test elements. New York, Teachers College, 1931.

Binet, A. La mesure en psychologie individuelle. Revue Philosophique, 1898, 46, 113-123.

Binet, A., & Simon, T. Methodes nouvelles pour le diagnostic scientifique des etats inferieurs de l'intelligence. L'Annee psychologique, 1905, 11, 163-190.

Binet, A., & Simon, T. Sur la necessite d'une methode applicable au diagnostic des arrierees militaires. Annales medico-psychologique, 1910.

Birnbaum, A. An efficient design and use of tests of a mental ability for various decision making problems. Series Report No. 58-16, USAF School of Aviation Medicine, Randolph, Texas, 1957.

Birnbaum, A. On the estimation of mental ability. Series Report No. 15, USAF School of Aviation Medicine, Randolph, Texas, 1958.

Bower, J. A history of western education. Civilization of Europe, sixth to sixteenth century, vol. 2. New York: St. Martin's Press, 1975.

Bright, O. T. Changes - wise and unwise - in grammar and high schools. In National Education Association Journal of Proceeding and Addresses, St. Paul, NEA, 1895.

Brown, W. Some experimental results in the correlation of mental abilities. British Journal of Psychology, 1910, 3, 296-322.

Brown, W., & Thompson, G. H. The essentials of mental measurement. Cambridge, Mass: Cambridge University Press, 1940.

Brownless, V. T., & Keats, J. A. A retest method of studying partial knowledge and other factors influencing item response. Psychometrika, 1958, 23, 67-73.

Burt, C. Experimental tests of general intelligence. British Journal of Psychology, 1909, 3, 94-177.

Cattell, J. M. Mental tests and measurements. Mind, 1890, 15, 373-381.

Cattell, R. B. Validity and reliability: A proposed more basic set of concepts. Journal of Educational Psychology, 1964, 55, 1-22.

Cochran, W. G. Early development of techniques in experimentation. In D. B. Owen (Ed.), On the history of statistics and probability. New York: Dekker, 1976.

Cremin, L. The transformation of the school. New York: Knopf, 1961.

Cronbach, L. J. Test "reliability": Its meaning and determination. Psychometrika, 1947, 12, 1-16.

Cronbach, L. J. Five decades of public controversy over mental testing. American Psychologist, 1975, 30, 1-14.

Cullen, M. J. The statistical movement in early Victorian Britain: The foundations of empirical social research. New York: Barnes & Noble, 1975.

DuBois, P. H. A test-dominated society: China, 1115 B.C.-1905 A.D. ETS Invitational conference on testing problems, Princeton: ETS, 1964.

DuBois, P. H. A history of psychological testing. Boston: Allyn and Bacon, 1970.

Edgeworth, F. Y. The element of chance in competitive examinations. Journal of the Royal Statistical Society, 1890, 53, 460-475, 644-673.

Englehart, M. D. Examinations. In W. S. Monroe (Ed.), Encyclopedia of educational research. New York: MacMillan, 1950.

Ferguson, G. A. Item selection by the constant process. Psychometrika, 1942, 7, 19-29.

Fisher, R. A. Statistical methods and scientific inference. New York: Hafner, 1956.

Freeman, F. N. Mental tests: Their history, principles and applications. Boston: Houghton Mifflin, 1926.

Graves, F. P. A history of education in modern times. New York: MacMillan, 1950.

Goodenough, F. L. A critical note on the use of the term 'reliability' in mental measurement. Journal of Educational Psychology, 1936, 27, 173-178.

Goodenough, F. L. Mental testing, its history, principles and applications. New York: Rinehart, 1949.

Guilford, J. P. Psychological measurement a hundred and twenty-five years later. Psychometrika, 1961, 26, 109-127.

Gulliksen, H. The content reliability of a test. Psychometrika, 1936, 1, 189-194.

Gulliksen, H. Measurement of learning and mental abilities. Psychometrika, 1961, 26, 93-107.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Guttman, L. A basis for scaling qualitative data. American Sociological Review, 1944, 9, 139-150.

Hambleton, R. K., & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Horst, A. P Item selection by means of maximizing function. Psychometrika, 1936, 1, 229-244.

Keats, J. A., & Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

Kelley, T. L. Interpretation of educational measurements. Yonkers-on-Hudson, New York, 1927.

Kelley, T. L., & Krey, A. C. Tests and measurements in the social sciences. Report of the Commission on the Social Studies, American Historical Association, Part IV. New York: Charles Scribner's Sons, 1934.

Kuder, G. F., & Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

Lawley, D. N. On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 1943, 61, Section A, 273-287.

Lazarsfeld, P. F. Latent structure analysis and test theory. In H. Gulliksen and S. Messick (Eds.), Psychological scaling: Theory and applications. New York: Wiley, 1960.

Lazarsfeld, P. F. The logical and mathematical foundations of latent struture analysis. In S. A. Stouffer, et al (Eds.), Measurement and prediction. Princeton: Princeton University Press, 1950.

Lentz, T. F., Hirshstein, B., & Finch, F. H. Evaluation of methods of evaluating test items. Journal of Educational Psychology, 1932, 23, 344-350.

Lincoln, E. A. The unreliability of reliability coefficients. Journal of Educational Psychology, 1932, 23, 11-14.

Lord, F. M. A theory of test scores. Psychometric Monographs, No. 7, 1952.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Marks, R. Providing for individual differences: A history of the intelligence testing movement in North America. Interchange, 1976-1977, 7, 3-16.

McCall, W. A. How to Measure in Education. New York: Macmillan, 1922.

Meitzen, A. History, theory, and technique of statistics. Philadelphia, 1891.

Meyer, A. E. Educational history of the western world. New York: McGraw Hill, 1965.

Monroe, W. S. Introduction to the theory of educational measurement. Boston: Houghton Mifflin, 1923.

Monroe, W. S. Educational measurement in 1920 and 1945. Journal of Educational Research, 1945, 38, 334-340.

Peterson, J. Early conceptions and tests of intelligence. Yonkers-on-Hudson, New York: World, 1925.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Neilsen & Lydiche, 1960.

Rice, J. M. Forum, 1897. Cited in W. H. Wilds & K. V. Lottich, Foundations of modern education. New York: Holt, Rinehart & Winston, 1970.

Ruch, G. M. The objective or new-type examination, an introduction to educational measurement. Chicago: Scott, Foresman, 1929.

Ruch, G. M., & deGraff, M. H. Corrections for chance and "guess" vs. "do not guess" instructions in multiple-response tests. Journal of Educational Psychology, 1926, 17, 368-375.

Rugg, H. O. Statistical methods applied to education. Boston: Houghton Mifflin, 1917.

Sharp, S. E. Individual psychology: A study in psychological method. American Journal of Psychology, 1899, 10, 329-391.

Spearman, C. Correlation calculated from faulty data. British Journal of Psychology, 1910, 3, 271-295.

Spearman, C. General intelligence objectively determined and measured. American Journal of Psychology, 1904, 15, 201-292.

Spring, J. H. Psychologists and the war: The meaning of intelligence and the Alpha and Beta tests. History of Educational Quarterly, 1972, 12, 3-15.

Strayer, G. D. Standards and tests for measuring the efficiency of schools or systems of schools. Bulletin, United States Bureau of Education, 1913, Whole No. 13: Report of the Committee of the National Council of Education.

Sylvester, D. W. Educational documents 800-1816. London: Methuen, 1970.

Thompson, G. O. B., & Sharp, S. History of mental testing. In T. Husen & N. Postlethwaite (Eds.), International encyclopedia of education: Research and studies, Oxford: Pergamon Press, 1983.

Thorndike, E. L. An introduction to the theory of mental and social measurements, 1904.

Thurstone, L. L. A method of scaling psychological and educational tests. Journal of Educational Psychology, 1925, 16, 433-451.

Thurstone, L. L. The reliability and validity of tests. Ann Arbor: Edwards, 1931.

Thurstone, L. L. The scoring of individual performance. Journal of Educational Psychology, 1926, 17, 446-457.

Thurstone, L. L. The unit of measurement in educational scales. Journal of Educational Psychology, 1927, 18, 505-524.

Toulouse, E., & Pieron, H. Technique de psychologie experimentale. Paris: Doin, 1904.

Tryon, R. C. Reliability and behavior domain validity: Reformulation and historical critique. Psychological Bulletin, 1957, 54, 229-249.

Tucker, L. R. Maximum validity of a test with equivalent items. Psychometrika, 1946, 11, 1-13.

Wilds, E. H., & Lottich, K. V. Foundations of modern education. New York: Holt, Rinehart & Winston, 1970.

Wissler, C. The correlation of mental and physical tests. Psychological Review, Monograph Supplement Vol. 8, No. 16, 1901.

Yerkes, R. M. (Ed.) Psychological examining in the United States Army. Memoirs of the National Academy of Sciences, 1921, 15, 1-890.