Handbook of Statistics, Vol. 26
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(06)26001-2

A History and Overview of Psychometrics

Lyle V. Jones and David Thissen[†]

1. Introduction

Psychometrics is defined well by the subheading that appeared under the title of the journal *Psychometrika* from its inception in 1936 until a cover redesign in 1984, "a journal devoted to the development of psychology as a quantitative rational science." Some definitions are even more inclusive; *WordNet* (developed by George Miller and his colleagues in the cognitive science laboratory at Princeton University; http://wordnet.princeton.edu/) defines *psychometrics* as "any branch of psychology concerned with psychological measurements."

In this chapter, we provide an alternative definition of psychometrics, eschewing the compactness of dictionary definitions, to focus on what researchers in the field of quantitative psychology do, and have done. For that, we begin with a beginning.

2. The origins of psychometrics (circa 1800–1960)

With the active support of graduate students and colleagues of L.L. Thurstone at The University of Chicago, the Psychometric Society was founded in 1935. The Society sponsored the journal *Psychometrika*, of which Volume 1, Number 1 appeared in March, 1936. At nearly the same time, the first edition of J.P. Guilford's (1936) *Psychometric Methods* was published. From one perspective, these events may serve to temporally locate the beginning of a formal sub-discipline of psychometrics. The founding of the Psychometric Society led to the publication in *Science* of Thurstone's presidential address to the meeting of the Society in September of 1936 that presented a strong plea to recognize a mathematical underpinning for psychological research (Thurstone, 1937).

Long before 1935, however, quantitative methods had achieved prominence in psychological research in Europe and in the United States. An early treatise on psychophysics, correlation, and ability testing is that of Brown and Thomson (1921), who had made important contributions to the development of test theory and factor analysis,

[†]We are grateful to Robert C. (Bud) MacCallum, Sandip Sinharay, Howard Wainer, and an anonymous reviewer for very helpful comments on an earlier draft of this chapter. Any errors that remain are, of course, our own.

Table 1
Some participants in various phases of the historical development of psychometrics

"Pioneers in psychometric methods" from Guilford's (1936) *Psychometric Methods*

Ernst Heinrich Weber	Edward Lee Thorndike
Gustav Theodor Fechner	Louis Leon Thurstone
Gerog Elias Müller	Alfred Binet
F.M. Urban	Truman L. Kelley
Sir Francis Galton	Lewis M. Terman
J. McKeen Cattell	Charles E. Spearman

Some of the participants in the WW I Army Alpha/Beta project

Lt. Col. Robert Yerkes	Lt. Arthur S. Otis
Major Lewis M. Terman	Edward L. Thorndike
Major Harold C. Bingham	L.L. Thurstone
Captain Edwin G. Boring	George M. Whipple
Lt. Carl C. Brigham	

Some of the participants in the WW II Army Air Force project[a]

Officers		Enlisted Personnel	
John C. Flanagan	Chester W. Harris	William Angoff	Joseph E. Mooney
Stuart W. Cook	Roger W. Heynes	Robert R. Blake	Albert Pepitone
Meredith P. Crawford	Nicholas R. Hobbs	Urie Bronfenbrenner	Harold M. Proshansky
Frederick B. Davis	Paul Horst	Benjamin Fruchter	Henry W. Riecken
Philip H. Dubois	Lloyd G. Humphreys	Nathaniel L. Gage	Milton Rokeach
Frank J. Dudek	John I. Lacey	Nathan M. Glaser	Elliot Steller
John E. French	Arthur W. Melton	Albert H. Hastorf	Rains B. Wallace
Robert M. Gagne	Neal E. Miller	Harold H. Kelly	Wilse B. Webb
Frank A. Geldard	William G. Molenkopf	Ardie Lubin	George R. Welch
Edwin E. Ghiselli	Donald E. Super	Samuel B. Lyerly	Joseph M. Wepman
James J. Gibson	John W. Thibaut	William McGrath	Wayne S. Zimmerman
J.P. Guilford	Robert L. Thorndike		

Some of the participants in the WW II OSS project[b]

Henry Murray	James G. Miller	Eliot Steller
Urie Bronfenbrenner	Donald W. McKinnon	Percival M. Symonds
Donald W. Fiske	O.H. Mowrer	Edward C. Tolman
John W. Gardner	Theodore M. Newcomb	Robert C. Tryon
David Krech	R. Nevitt Sanford	

[a]A complete listing of about 1500 names associated with the Army Air Force project is provided by DuBois (1947, pp. 377–392).
[b]A complete listing of the staff associated with the OSS project is provided in the Office of Strategic Services Assessment Staff (1948) report.

respectively. Other "pioneers in psychometric methods" were featured in portraits in the frontispiece of the first edition of Guilford's (1936) *Psychometric Methods*; they are listed here in the top panel of Table 1.

Figure 1 places some of the important figures in the early history of psychometrics on a timeline and illustrates some of the linkages among them. It is of interest to trace

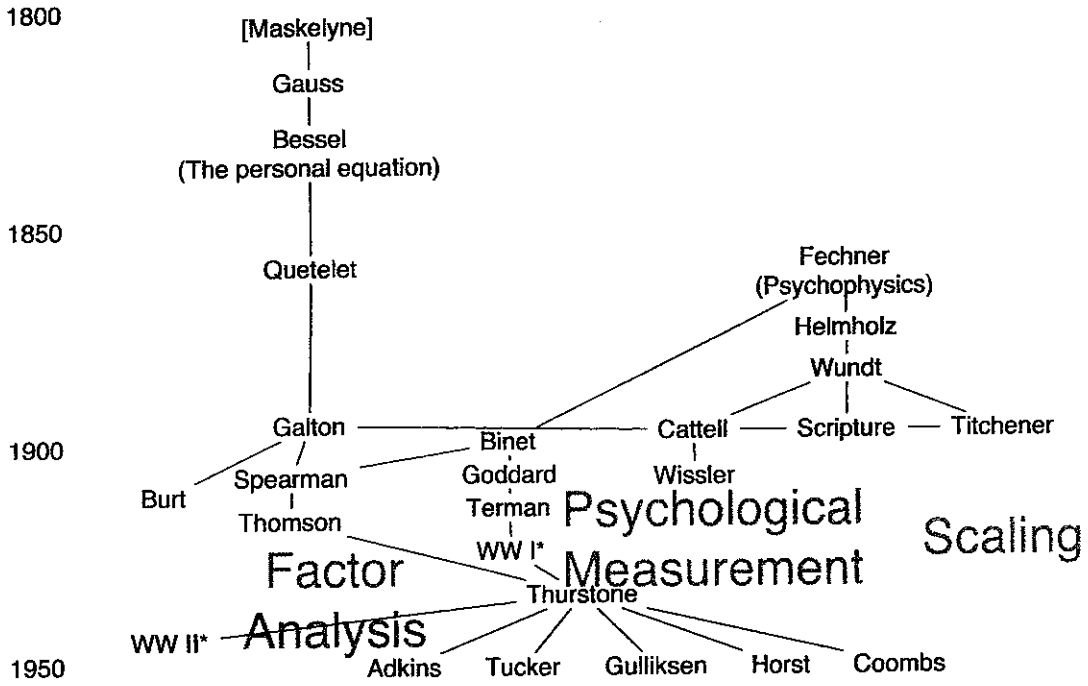

Fig. 1. Some of the important figures in the early history of psychometrics on an approximate timeline.
*Some of the participants in the World War I and World War II projects are listed in Table 1.

the influence of some of those scholars and others associated with them, for their works form the roots of contemporary psychometrics.[1]

2.1. *Psychophysics*

The founding of psychometrics might well be thought to correspond with the founding of academic psychology, marked by the publication in 1860 of physicist-philosopher Gustav Fechner's *Elemente der Psychophysik*. Fechner defined psychophysics as "an exact science of the fundamental relations of dependency between body and mind" (Gulliksen in the foreward to Torgerson, 1958, p. v). Boring (1929, p. 286) wrote that Fechner "set experimental quantitative psychology off upon the course which it has followed." Fechner experimentally established the lawful relation between the measure of a physical stimulus (R) and the measure of an observer's sensation thereof (S) as

$$S = k \log R,$$

a formulation that stimulated psychometric research not only in Europe, especially in Germany, but somewhat later in the United States as well.

Hermann von Helmholtz, a physicist–physiologist first at Bonn, then Heidelberg and finally Berlin, contributed to many scientific fields, including the psychology of sensation. He accepted Fechner's proposed logarithmic relation between sensation and stimulus, and he adopted psychophysical methods to support his theories of vision,

[1] The informative review provided by Hilgard (1987) of contributions of individual psychologists has contributed in a variety of ways to the following discussion.

hearing, and tactile sensation. Helmholtz' former assistant was Wilhelm Wundt, whose Leipzig laboratory in 1879 became the first psychology laboratory in the world. Wundt and his students also employed psychophysical methods as primary tools of experimental psychology. In the United States, too, psychophysics was the dominant experimental method for psychology, imported from Germany by psychologists James McKeen Cattell (the first American PhD of Wundt in 1886) at Pennsylvania, Edward Scripture (a Wundt PhD in 1891) at Yale, and Edward Titchener (a Wundt PhD in 1892) at Cornell, among others.

2.2. *Early testing of individual differences*

From another perspective, psychometrics had been founded in the early years of the 19th century by two astronomers, Friedrich Wilhelm Bessel and Carl Friedrich Gauss.

In 1796, Nevil Maskelyne, astronomer royal at the Greenwich Observatory, had dismissed his assistant David Kinnebrook because Kinnebrook reported times of stellar transit nearly a second later than did Maskelyne. The incident was recorded in *Astronomical Observations at Greenwich Observatory*, and was mentioned in a history of the observatory that appeared in 1816 that was noticed by Bessel, director of a new observatory at Königsberg. Bessel acquired and studied the earlier notes of Maskelyne, then undertook a series of independent studies at Königsberg. In 1823 Bessel reported average differences of 1.0 and 1.2 seconds between his observations and those of two other astronomers, a finding that astonished other astronomers who had expected far smaller individual differences.

Bessel also became aware of earlier work by Gauss, an astronomer at Göttingen, who in 1809 had presented the theory of errors of observation employing a statistical distribution now known as the normal or the Gaussian distribution. Bessel then developed a "personal equation" to correct observations for differences among observers. He also discovered that individual differences became smaller with the use of a chronometer with half-second rather than full-second beats. As a result of his studies, the accuracy of astronomic observations improved markedly (see Boring, 1929, pp. 134–142). Bessel's work anticipated the reaction-time experiments that continue to be popular in psychological laboratories.

In Great Britain, Frances Galton, motivated by his interest in human eugenics, was a champion of the measurement of individual differences. Galton was influenced by Adolph Quetelet, Belgian statistician, who had applied the Gaussian law of error to a wide range of human data. Galton designed apparatus to measure a variety of bodily dimensions. Galton's Anthropometric Laboratory opened in London in 1884, and collected data on about 10,000 persons over the ensuing six years. A prominent supporter of Galton's program in the United States was Joseph Jastrow, from 1888 a professor of psychology at the University of Wisconsin. Jastrow had earned a PhD with G. Stanley Hall at Johns Hopkins in 1886. He began a correspondence with Galton in 1887. By 1893, Jastrow (together with anthropologist Franz Boaz) established a version of Galton's Anthropometric Laboratory at the World's Columbian Exposition in Chicago. Galton visited the Laboratory, as did thousands of others whose physical features were measured.

During the late 1880s, James McKeen Cattell spent time in Galton's laboratory and developed an interest in psychometric measures. In 1889, Cattell returned from England to a psychology professorship at the University of Pennsylvania where he established a laboratory of psychology. Largely ignoring the bodily measurements that had dominated Galton's program, he measured his students' hand strength, rate of movement, pain threshold, reaction times, judgments of elapsed time, etc., referring to these as "mental tests" (Cattell, 1890) and expecting them to be related to other evidence of mental ability (Sokal, 1982, p. 329). Cattell moved to Columbia University in 1891, where he was permitted by the university administration to record for all entering freshman students the results from his mental testing program. He promoted his testing program, but without setting forth either a theoretical or an operational justification for it.

The value of the sensory and motor tests developed by Cattell and others was challenged by Clark Wissler, a graduate student at Columbia who studied both with Cattell and with anthropologist Boaz. Wissler found that, for each of Cattell's "mental tests," the correlations with class grades were essentially zero (Wissler, 1901). That finding effectively ended the testing of sensory reactions and motor skills for assessing mental functioning, although the utility of such tests for evaluating sensory disorders in hearing, vision, and tactile sensitivity continues to be recognized.

2.3. *Psychological testing and factor analysis*

At about the same time that anthropometric testing was gaining public attention in Great Britain and the U.S., Alfred Binet and Théophile Simon in France were developing an alternative approach, specifically designed to assess higher cognitive abilities. Binet and Simon (1905) introduced the concept of mental age, determined by earning a test score that is characteristic of children of a given chronological age, and they developed tests suitable for estimating a child's mental age. The intelligence testing movement also was given a boost by the work of Charles Spearman in Great Britain, who analyzed results from a battery of cognitive tests and interpreted the analyses as providing support for a general intelligence factor, g, common to all of the tests (Spearman, 1904).

In 1906, Henry H. Goddard became research director at the Vineland (NJ) Training School for the Feeble-Minded. By 1908 he had visited France, and he provided a translation of the first version of the Binet–Simon scale (Goddard, 1908). Other translations followed; the best known was that of Stanford University psychology professor Lewis M. Terman (1916). Terman adopted the proposal of William Stern, psychologist at the University of Breslau, to construct a quotient of mental age divided by chronological age. Terman multiplied this quotient by 100 to define an IQ score, based upon a child's responses to items on the Stanford–Binet.

In April of 1917, at the initiative of Robert M. Yerkes, president of the American Psychological Association, the APA Council voted "that the president be instructed to appoint committees from the membership of APA to render to the Government of the United States all possible assistance in connection with psychological problems arising in the military emergency" (Yerkes, 1921, p. 8). Almost immediately, this effort was joined by the National Research Council, that proposed to the Surgeon General of the Army a plan for the psychological examination of all Army recruits. By August, 1917, the plan was approved by the Secretary of War.

During the initial planning period in July, 1917, a statistical unit was formed, with Edward L. Thorndike as statistician and Arthur S. Otis and L.L. Thurstone his assistants. Later Major Robert Yerkes acted for the Surgeon General as head of psychological work, assisted by many others, including those listed in the second panel of Table 1. Their efforts culminated in the establishment of a School for Military Psychology at Camp Oglethorpe, Georgia, the development of batteries of psychological tests (Army Alpha and Beta), and the testing during 1918 of nearly two million U.S. Army recruits.

Arthur S. Otis, as a Stanford graduate student, had worked with Terman. For his dissertation Otis devised procedures for scoring multiple-choice items that made feasible group rather than individual testing; these procedures saw their first large-scale application with the administration of the Army tests. Subsequently, Otis's *Group Intelligence Scale*, modeled after the Army Alpha, was the first commercial group-administered mental test. It was published by the World Book Company in 1918. Terman also published a very slightly modified version of Army Alpha as the *Terman Group Test of Mental Ability* in 1920, again with the World Book Company. Otis joined the World Book staff in 1921. In 1923, Terman and colleagues produced the *Stanford Achievement Test*, an objective measure of what students had learned in school.

After noting the impact in the 1920s of Freud's sexual theories on public awareness of psychology, one respected historian noted that "Even more important in popularizing psychology were the Army 'intelligence' tests and the debates they aroused" (Leuchtenburg, 1958, p. 164). Some effects of the testing controversy on educational reform are presented by Cremin (1961, pp. 185–192).

In the early 1920s, Louis Leon Thurstone served as chair of the Department of Applied Psychology at the Carnegie Institute of Technology. In 1923–1924, Thurstone was supported by the Carnegie Corporation of New York to spend a year in Washington, DC to devise improvements to the federal civil service examinations used in the selection of personnel for U.S. government employment. While fulfilling that mission, Thurstone also prepared for the American Council on Education (ACE) a *Psychological Examination for High School Graduates and College Freshmen*. That exam produced two scores, a linguistic score, "L" and a quantitative score, "Q" (Thurstone, 1924). In 1924 Thurstone joined the faculty at The University of Chicago where he and his wife, Thelma Gwinn Thurstone, continued to produce annual versions of that test until 1947. In that year, the Educational Testing Service (ETS) was founded, drawing together the testing programs of ACE, the Carnegie Foundation for the Advancement of Teaching, and the College Entrance Examination Board (CEEB, now known as the College Board). The Scholastic Aptitude Test (SAT) produced by ETS also provided two scores, one for verbal aptitude that correlated higher with college grades in English and history, and one for mathematical aptitude that correlated higher with grades in mathematics and engineering.

During the years in which the Thurstones were developing the ACE tests, L.L. Thurstone also was devising methods for multiple factor analysis (Thurstone, 1931a, 1935). This led to the collection of results from a battery of tests, first from freshmen at the University of Chicago, then from Chicago public high school students. Factor analyses of those data identified seven to nine separate factors of intellect (Thurstone, 1938), results that can be set in opposition to the Spearman one-factor g theory. (Later, when he

allowed for factors to be correlated, Thurstone granted the possibility that Spearman's *g* might be found as a single second-order factor, the result of analyzing the correlations among the primary factors.)

Applications as well as theoretical developments in factor analysis were stimulated by the appearance of Thurstone's authoritative book *Multiple Factor Analysis* (Thurstone, 1947). Mulaik (1986) provides a review of research literature relating to factor analysis for the 50 years from the 1930s to the 1980s.

In 1941, prior to the entry of the U.S. into World War II, John Flanagan played a role for psychologists similar to that of Yerkes just 24 years earlier. The War Department approved Flanagan's proposal to launch an ambitious testing program within the Army Air Corps to establish the validity of a test battery for selecting and classifying applicants for Air Corps flight crews, initially for pilots, navigators, and bombardiers. A guide to the 19 detailed book-length reports of the program is given by Flanagan (1948). Some 1500 officers and enlisted personnel were engaged in the effort, including many with names that would subsequently become familiar; some are listed in the third panel of Table 1. Scores of the Air Force psychologists became distinguished as contributors to research psychology in post-war years. The Army Air Force's project entailed a unique validity study in which all applicants were accepted in the program, although each had been required to complete a battery of tests to be used potentially for selection and classification. By relating test scores to success in training, a final battery was validated and used later to screen subsequent applicants (DuBois, 1947; Guilford and Lacey, 1947).

After the war, John Flanagan founded the American Institutes for Research (AIR), for the initial sole purpose of providing to commercial airlines a set of selection procedures highly similar to those developed in the Army program. More recently, AIR has engaged in a wide range of research activities, many of which contribute to governmental efforts in support of educational assessment. Other participants in the Army Air Force's program returned to quantitative, experimental, or social psychology programs at a wide range of universities, as well as testing agencies.

In another effort shrouded in a great deal more secrecy during the second World War, a collection of psychologists developed and implemented an extraordinary assessment system for the selection of personnel (otherwise known as spies) for the Office of Strategic Services (OSS). This work is described in a military report, the declassified version of which was published under the titles *Assessment of men: Selection of Personnel for the Office of Strategic Services* (Office of Strategic Services Assessment Staff, 1948) and *Selection of Personnel for Clandestine Operations: Assessment of Men* (Fiske et al., 1993). The staff of "Station S" (for "secret") was largely organized by Henry Murray; some of the dozens of staff members are listed in the final panel of Table 1. The assessment included paper-and-pencil instruments and projective tests as well as a large number of performance and situational exercises that were rated by observers. Extensive data analysis, including the (then new) techniques of multiple correlation and factor analysis, supported the clinical judgment of the staff in recommending candidates. While the extravagance of the multi-day OSS assessment has had few, if any, equals since World War II, the experiences of the OSS staff in that environment were to have extensive consequences for research in personality measurement and personnel selection for a generation of psychologists (Handler, 2001).

Over the decades following World War II, numerous school achievement tests were developed to compete with Terman's Stanford Achievement Test. A number of for-profit and non-profit companies were in the business of producing achievement tests, among them the Psychological Corporation, the Iowa Testing Program, Science Research Associates, Educational Testing Service, CTB/McGraw-Hill, and others. The National Assessment of Educational Progress, a program to monitor school achievement for the nation, was launched by Ralph W. Tyler in the 1960s and periodically has monitored achievement levels of national samples of school children (and state samples, since 1990) to provide "the nation's report card" (see Jones and Olkin, 2004). More recently, almost all states have adopted achievement tests for students in public schools; a mandate for such tests is a feature of the federal program "No Child Left Behind."

2.4. Organizational developments

In 1935, as noted earlier, the Psychometric Society was founded (its charter president was L.L. Thurstone) and the journal *Psychometrika* was launched. In the following year J.P. Guilford (1936) published *Psychometric Methods* that became a standard college text for the study of psychological measurement. The journal *Educational and Psychological Measurement* appeared in 1941. In 1946, the American Psychological Association (APA) created a Division of Evaluation and Measurement, and L.L. Thurstone was its first president. (The name of the division was changed in 1988 to Division of Evaluation, Measurement, and Statistics.) The Society of Multivariate Experimental Psychology was founded in 1960 (according to its website, http://www.smep.org) as an organization of researchers interested in multivariate quantitative methods and their application to substantive problems in psychology; that society publishes the journal *Multivariate Behavioral Research*. In 1964, the National Council on Measurement in Education inaugurated the *Journal of Educational Measurement*, publishing articles on psychometrics as applied to measurement in educational settings. The Society for Mathematical Psychology began to meet in 1968, and publishes the *Journal of Mathematical Psychology*. A number of other journals, including *Applied Psychological Measurement*, *Psychological Methods*, the *Journal of Educational and Behavioral Statistics*, the *British Journal of Mathematical and Statistical Psychology*, *Structural Equation Modeling*, and *Behaviormetrics* have joined the set that publishes psychometric research.

From these historical roots, and with the infrastructure provided by emerging scientific societies, by the second half of the 20th century psychometrics had staked out its intellectual turf: psychological scaling, educational and psychological measurement, and factor analysis. Extensions of those three core topics, along with numerous contributions to applied statistics motivated by the data collection in empirical psychology, have defined the field through the second half of the last century and into the 21st.

3. Psychological scaling

3.1. Thurstone's scaling models and their extensions

S.S. Stevens (1951) wrote that "Measurement is the assignment of numerals to objects or events according to rules." When measurement, thus broadly defined, makes use of

behavioral or psychological data, it is called *psychological scaling*, because the assignment of numerals places the objects or events on a *scale*. To make use of behavioral or psychological data, the "rules" for the assignment of numerals are usually based on mathematical or statistical models for those data. Psychological scaling had its origins in psychophysics, with the birth of psychology as a science; but it has expanded to be used in many domains.

From its beginnings in the 19th century, psychophysics had continued to enjoy support in academic psychology during the mid-1920s and early 1930s, a period in which L.L. Thurstone published a remarkable set of papers that proposed adapting psychophysical methods to scale psychological functions that had no clear physical correlates, e.g., human abilities, attitudes, preferences, etc. The Thurstonian scaling methods were generated directly from employing the Gaussian distribution, first to repeated responses from a given individual, then with an extension to the distribution of responses for many individuals. These scaling methods almost immediately were adopted as measurement tools for research in social psychology, and thereby became an important component both of academic and applied psychology. Thurstone considered his development of scaling theory and practice to have been his most important contribution. An integrative summary of Thurstonian scaling procedures is that of Bock and Jones (1968). As noted in subsequent sections of this chapter, Thurstone's scaling methods stimulated later developments in multidimensional scaling and in item response theory. They also influenced the formulation of related scaling methods such as Luce's choice model (Luce, 1959).

In 1950, the Social Science Research Council appointed a Committee on Scaling Theory and Methods to review the status of scaling procedures and their relation to basic research in the social sciences. Committee members were Paul Horst, John Karlin, Paul Lazersfeld, Henry Margenau, Frederick Mosteller, John Volkmann, and Harold Gulliksen, chair. The committee engaged Warren Torgerson as a Research Associate to review and summarize the literature on psychological scaling. The result was the publication of a book that stood for many years as the authoritative summary of scaling procedures (Torgerson, 1958).

3.2. *Multidimensional scaling*

Various strategies for *multidimensional scaling* (MDS) borrow from Thurstone's scaling models (and their descendents) the idea that similarity data (as might be obtained using various experimental procedures and human judgment) can be represented spatially. However, where Thurstone's scaling models represent objects with real numbers on a single dimension, MDS most often represents objects as points in two- (or higher) dimensional space. In the course of this extension of dimensionality, most procedures for MDS are more data analytic in character, and less based on any psychological process model, than were Thurstone's original procedures. As such, MDS procedures form a bridge between the original methods of psychological scaling and a number of purely data-analytic procedures that are discussed in section 6 of this chapter.

Torgerson (1952) introduced multidimensional scaling using "metric" algorithms that treat the numerical values of the similarity data as if they are on an interval scale.

In the 1960s, simultaneous advances in the theory of statistical optimization and computing made possible the implementation of algorithms for "nonmetric" MDS, which seeks to derive a spatial representation of objects, with coordinates on an interval scale, making use only of the ordinal properties of the original similarity data. A theoretical paper by Luce and Tukey (1964) together with algorithms and computer programs by Shepard (1962a, 1962b), Kruskal (1964a, 1964b), Guttman and Lingoes (Lingoes, 1965), and Young and Torgerson (1967) provided the basis for countless studies in the fields of sensation, perception, and cognition in the 1960s and 1970s (see, for examples, Young and Hamer, 1987). The properties of nonmetric MDS, and additional improvements to methods for its implementation, were among the most prominent topics of psychometric research between the mid-1960s and the mid-1980s.

Tucker and Messick (1963) originated the idea that individual differences could be modeled in MDS, by modeling separate perceptual spaces for individuals having different "viewpoints" about stimulus interrelationships. Additional procedures that combined the idea of individual differences with a spatial representation underlying judgments of similarity were among the new procedures introduced around 1970 (Horan, 1969; Carroll and Chang, 1970; Tucker, 1972; Bloxom, 1974). Torgerson (1986) provided a critical summary of this most active period in the development of procedures for psychological scaling.

While basic MDS has subsequently become a standard data-analytic technique, continuing developments that extend these ideas remain active topics of research, some of which are mentioned in Section 6.

4. Psychological measurement (test theory)

4.1. True score theory

A number of psychological testing programs were initiated in the first half of the 20th century, but were launched with little attention to an underlying theory of testing. When L.L. Thurstone joined the faculty of the University of Chicago, he immediately designed a curriculum for a course in test theory. Later, the material was published under the title, *The Reliability and Validity of Tests* (Thurstone, 1931b). For several decades, this work served to foster numerous contributions to the psychometric literature, largely refinements and embellishments to traditional test theory based on the concept of the "true score."

Test theory is the archetype of a problem unique to psychological research that requires a statistical solution. Considering a test score from a statistical point of view, it is highly desirable to derive an ancillary statement of its precision. In the most basic approach to true score theory, the test score X is considered the sum of a true score T and a random error E,

$$X = T + E.$$

The standard deviation of the errors E is a statement of the (lack of) precision, or standard error, of the test score. That standard error can be estimated using an estimate of the reliability of the score X, where the reliability is the squared correlation between observed scores and the true scores.

Procedures to estimate reliability were developed over the first half of the 20th century, beginning with the seminal papers by Spearman (1910) and Brown (1910) that established the algebraic relation between test length and reliability for (hypothetical) parallel items. The first widely-used internal-consistency estimates of reliability (permitting reliability to be estimated from a single administration of a test, without some arbitrary division into halves) came from Kuder and Richardson (1937) famous paper, in which the equations numbered 20 and 21 came to be the names of statistics estimating reliability (KR-20 and KR-21). Hoyt (1941) provided a more general statement of internal consistency reliability, which was popularized by Cronbach (1951) as "coefficient α." Gulliksen's (1950) landmark text, *Theory of Mental Tests*, provided a summary of the research and developments of the first half of the 20th century, and was the standard theoretical and procedural guide for testing until the publication of the elegant axiomatic treatment of true score theory presented by Lord and Novick (1968). Variants of true score theory, sometimes fairly complex, continue to be used to provide quantitative statements of precision of test scores; Feldt and Brennan (1989) summarize the underlying theory and procedures for tests with a variety of structures.

Cronbach et al. (1972) expanded the concept of reliability to become *generalizability* over a number of possible facets of test administration, for example items or tasks, occasions of measurement, and judges (for responses scored by raters). Generalizability theory has become extremely important for the evaluation of "authentic" or "performance" assessments that often involve extended tasks or exercises administered over a period of time with responses evaluated by trained raters. Brennan (2001) summarizes continuing developments in this area.

During the latter half of the 20th century, as enriched notions of generalizability evolved, there emerged also broadened views of test validity. Cronbach and Meehl (1955) presented the first detailed exposition of "construct validity" as an adjunct to traditional predictive validity. Samuel Messick, in a series of papers culminating with Messick (1980), set forth a further set of considerations that clarify the distinction between construct validity, which he considered *the* validity of a test, and other features such as predictive utility, content coverage, and so on. Messick's formulations have been influential on subsequent revisions of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999).

Instruction in test theory is often the only presence of quantitative psychology in the undergraduate psychology curriculum aside from introductory statistics and research methods courses. Influential textbooks on testing include the well-known texts by Anastasi (most recently, Anastasi and Urbina, 1996) and Cronbach (1990) among many others. Summaries of the literature on test theory can be found for the period 1936–1960 in an article by Gulliksen (1961) and for 1961–1985 in a similar article by Lewis (1986).

4.2. Item response theory (IRT)

A global shift in graduate instruction and research in test theory accompanied the publication in 1968 of Frederic Lord and Melvin Novick's *Statistical Theories of Mental Test Scores* (Lord and Novick, 1968). In addition to a thorough, modern mathematical and statistical codification of true score theory, that volume included the first comprehensive coverage of IRT. Along with chapters by Lord and Novick on the normal ogive model, additional chapters by Allen Birnbaum (1968) introduced the logistic model, extended an idea that had been suggested by Lord (1953b) to account for guessing in what has come to be called the *three-parameter logistic* model (3PL), and advocated the use of information functions in test development.

IRT focuses the statistical analysis of test data on the responses to items, rather than on the total summed test score which was the basis of true score theory. Most of the models and methods that fall within IRT assume that one or more unobserved (or latent) variables underlie the responses to test items in the sense that variation among individuals on those latent variables explains the observed covariation among item responses (Green, 1954). In IRT models, the relation between the position of individuals on the latent variable(s) and the item responses is described by statistical models that describe the probability of an item response as a function of the latent variable(s). Examples of such models are the aforementioned normal ogive and logistic models; Bock and Moustaki (this volume) provide a detailed overview of IRT, including descriptions of many other models.

Before 1968, research in IRT had proceeded very slowly, with the occasional theoretical contribution, and virtually no application. The seeds of IRT were planted in a seminal paper by L.L. Thurstone (1925) in which he used ideas closely related to what would become the normal ogive model to describe Burt's (1922) data obtained with a British translation of Binet's scale, and another paper by Symonds (1929) in which the idea of the normal ogive model was used to describe data obtained from spelling tests. Guilford's (1936) *Psychometric Methods* described analysis of test data using IRT as though it might soon provide a practical method of item analysis; it actually required more than 40 years for that prediction to be fulfilled.

In a relatively obscure paper published in the *Proceedings of the Royal Society of Edinburgh*, Lawley (1943) described the application of the (then relatively new) method of maximum likelihood to the estimation of ability using the normal ogive model. However, because no practical method existed to produce the requisite item parameters for the model, Lawley's statistical development had no practical use. Tucker (1946) used the normal ogive model as a theoretical device to show that a maximally reliable test is made up of items with difficulty equal to the average ability of the examinees, and Lord (1952, 1953a) also used IRT as a theoretical basis for a description of the relation between ability and summed scores. The work by Lord (1952) and (intellectually) closely related work by the mathematical sociologist Paul Lazarsfeld (1950) clarified the nature of IRT models and emphasized the idea that they were "latent variable" models that explain the observed covariation among item responses as accounted for by an underlying, unobserved variable related to each response. B.F. Green's (1954) chapter in the *Handbook of Social Psychology* provides a lucid explanation of the theoretical struc-

ture of IRT models and the related latent class models that assume *local independence*: independence of item responses conditional on the value of the latent variable.

Lord and Novick's (1968) volume placed the normal ogive IRT model on a sound theoretical basis as an integration of a latent variable representing individual differences with the model Thurstone (1927) introduced as the "law of comparative judgment." After 1968, research and development in item response theory increased extraordinarily rapidly. The problem of item parameter estimation was brought into sharp statistical focus by Bock and Lieberman (1970), and a practical method based on maximum likelihood was described by Bock and Aitkin (1981). The latter method is used by most general purpose IRT software. Samejima (1969, 1997) provided the ground-breaking suggestion that IRT could be extended to the analysis of items with more than two ordered responses; Bock's (1972, 1997a) model for more than two nominal responses quickly followed. A number of refinements of, and alternatives to, these models have since been described (see Thissen and Steinberg, 1986). In the 1990s, following a seminal article by Albert (1992), research on applying Markov chain Monte Carlo (MCMC) techniques to estimation in item analysis went forward in several laboratories around the world, opening the way for more complex models (in which IRT plays a part) than are likely feasible using a maximum likelihood approach (Béguin and Glas, 2001; Bradlow et al., 1999; Fox and Glas, 2001, 2003; Patz and Junker, 1999a, 1999b; Wang et al., 2002).

In original work in Denmark that was largely independent of the American and British research that had gone before,[2] Georg Rasch (1960) developed an item response model by analogy with the properties of physical measurement. While Rasch's "one-parameter" logistic model (so named for its only item parameter that reflects difficulty) is closely related to Birnbaum's (1968) two-parameter logistic model, Rasch (1961, 1966, 1977) subsequently developed a philosophical basis for his model that established the Rasch tradition as an alternative to the models of the Thurstone–Lord–Birnbaum traditions. Subsequently, the "Rasch family" (Masters and Wright, 1984) has also expanded to include generalized models for items with more than two response alternatives, and a number of specialized models as well.

Lewis (1986) summarizes the literature on test theory for the period of IRT's explosive growth in the 1970s and early 1980s, and Bock (1997b) provides a more complete history of IRT. Perusal of the contents of this volume makes clear that IRT remains one of the most active areas of research in psychometrics.

4.3. *Equating, test assembly, and computerized adaptive testing*

The problem of *test equating* arises in its most commonly recognized form from the fact that large-scale assessments (such as the SAT) make routine use of alternate forms (certainly over time, and sometimes simultaneously), but all forms must yield comparable reported scores. Numerous statistical algorithms have been proposed for various data

[2] Rasch (1960) refers to Lord (1953b) only once, in a footnote on p. 116, and the normal ogive model only briefly there and on pp. 118 and 122. Rasch also refers, in another footnote, to the choice model (Bradley and Terry, 1952; Luce, 1959); however it appears that these footnotes were added after the fact of Rasch's model development.

collection designs to provide equated scores with various properties, and book-length treatments of the subject describe those methods (Holland and Rubin, 1982; Kolen and Brennan, 2004; von Davier et al., 2003), as do chapters in each edition of *Educational Measurement* (Flanagan, 1951; Angoff, 1971; Petersen et al., 1989; Holland and Dorans, in press). Holland and Dorans (in press) trace the origins of test equating to work by Truman Kelley (1914); Kelley's (1923) *Statistical Methods* included a chapter on "comparable measures" that describes linear equating and equipercentile equating, and refers to alternative proposals by Otis (1916, 1918). For most of the 20th century, test equating was a highly specialized technical process of interest primarily to its practitioners at institutions that created large-scale assessments with alternate forms. While a large psychometric literature accrued on this topic, it was not highly visible.

However, in the 1990s the increasingly salient (and burdensome) use of educational achievement tests in the United States led to requests to link scores on different tests, with the ultimate goal of using the results obtained from one assessment as substitute scores for those that would have been obtained on another test, to reduce the amount of time spent on achievement testing by having the results from one test serve double duty. Test linking then became the subject of acts of Congress, and two committees of the National Research Council published reports describing the many reasons why scores on distinct tests cannot, in general, be made comparable in all respects by any statistical processes (Feuer et al., 1999; Koretz et al., 1999).

Based on a substantial body of work inspired by attempts in the 1990s to extend the reach of test equating beyond its previous scope, Holland and Dorans (in press) use test *linking* as a generic term to encompass (i) *equating*, for producing interchangeable scores, (ii) *scale aligning*, for producing comparable scores (in some senses), and (iii) *predicting*, to make the best prediction of the score on one test from the score on another. For each of those three specific goals, many statistical techniques have been employed, including procedures based on summed scores and on IRT.

Test linking remains an active area of psychometric research as evidenced by the fact that two journals recently published special issues devoted to that area: The *Journal of Educational Measurement*'s special issue "Assessing the population sensitivity of equating functions" appeared in Spring, 2004, and *Applied Psychological Measurement*'s special issue on "Concordance" appeared in July, 2004.

Approaches to test linking may be based on true score theory or on IRT. However, two topics of currently active research in test theory rely primarily on IRT: computerized test assembly and computerized adaptive testing (CAT).The goal of computerized test assembly is usually to create parallel test forms, or at least alternate forms of a test that may be scored on the same scale, from a collection of items (a "pool") that has usually been calibrated using IRT. The scale of the problem can be very large: Hundreds of items may be cross-classified by dozens of content- and format-based attributes as well as by their statistical properties as described by IRT parameters. The numbers of combinations and permutations of those items that are possible if they are to be distributed onto fairly long tests is (literally) astronomical. Human test assembly specialists have assembled test forms for decades using various unspecified heuristic algorithms; computerized test assembly algorithms are intended to assist the human test assemblers (or perhaps replace them) with procedures that produce alternate test forms very quickly (often in ways that

will not require subsequent equating of the alternate forms). A special issue of *Applied Psychological Measurement* (1998, volume 22, number 3) includes articles describing several computerized test assembly procedures (see van der Linden, 1998).

Perhaps the most extreme examples of computer-generated tests with linked score scales are those administered by CATsystems, in which each respondent may see a different set of items, sequentially selected to provide optimal information at each stage of the testing process. Volumes edited by Wainer et al. (2000) and van der Linden and Glas (2000) provide accessible entry points for (roughly) three decades of extensive research and development on the subject of CAT. Much of the modern development of IRT in the 1970s and early 1980s was supported by the U.S. Department of Defense in the project to create a CAT version of the Armed Services Vocational Aptitude Battery, or ASVAB. While widespread participation of academic quantitative psychologists in that project has ended and the test is operational, the effects of the project on large-scale testing persist.

5. Factor analysis

5.1. Factor analysis and rotation

Factor analysis is a statistical solution to a psychological problem: Almost as soon as (modern) psychological tests were invented, it was observed that the scores on tests of distinct aspects of ability were correlated across persons. Spearman (1904) sought to explain that correlation by appeal to the idea that the scores on all of the tests reflected, in some measure, variation on an underlying variable he called g (for general intelligence); that article is usually considered the beginning of factor analysis.

While subsequent British research began to consider the idea that there might be more than one factor of ability, it was Thurstone's (1938) *Psychometric Monograph* on the *Primary Mental Abilities* that initiated a line of factor analytic research on the structure of the intellect that would continue for decades. There were problems determining the number of factors, and then of estimating the parameters—the factor loadings, which are the regression coefficients of the observed test scores on the unobserved factors, and the unique (error) variances. Factor analysis was invented long before the theoretical statistics had been developed that would provide methods for the estimation problem for a model with so many parameters, and also long before computational power was available to solve the problem (had suitable algorithms been available). Before the 1960s, creative heuristic methods, largely due to Thurstone, were used.

While others had managed maximum likelihood (ML) estimation for the factor analysis model earlier, it was Jöreskog's (1967) algorithm, and Jennrich and Robinson's (1969) alternative, coinciding with the development of sufficiently fast electronic computers, that gave rise to practical general computer implementations of a statistically optimal procedure. While the earlier heuristic algorithms were also implemented in standard statistical software, the ML procedures permitted consideration of *Factor Analysis as a Statistical Method*, the title of Lawley and Maxwell's (1963, 1971) book.

Another daunting problem with factor analysis came to be called the rotational indeterminacy of factor-analytic solutions. For any factor-analytic model with more than

one common factor, there are an infinite number of linear transformations that yield the same fit to the data; in one graphical way of displaying the results, these transformations may be obtained by rotating the axes of the graph (hence the name). Thurstone (1947) proposed the criterion of "simple structure" to select an interpretable version of the solution. For the first decades factor analysts used time-consuming, labor-intensive graphical methods to rotate the solution to simple structure. Before factor analysis could be used routinely, some analytical/computational method for rotation was required. Mulaik (1986, p. 26) wrote that "solving the rotation problem became a kind of Holy Grail for factor analysts to pursue, and many a factor analyst was to make his reputation with a workable analytic scheme of factor rotation." Horst (1941) proposed one of the first analytic methods; Kaiser's (1958) "varimax" method is likely the most successful of the procedures for orthogonal factor solutions. "Direct oblimin" (Jennrich and Sampson, 1966), a modification of a procedure proposed by Carroll (1953), is an example of an analytical rotation algorithm for oblique factor solutions. Browne (2001) provides an excellent overview and review of analytic rotation methods.

Between the 1930s and the 1960s, factor analysis was basically the exclusive domain of quantitative psychologists. By 1976, a textbook in factor analysis was a very large volume indeed (see Harmon, 1976); much had been accomplished. With computer applications prepared to perform the computations involved in the statistical estimation and analytical rotation of the factor loadings, after the 1970s factor analysis became a generally available tool for research psychologists.

John B. Carroll's (1993) 800-page volume *Human Cognitive Abilities: A Survey of Factor Analytic Studies* may represent the culmination of efforts extended toward the original purpose of factor analysis – to draw inferences about the trait structure of human cognition from correlations among test scores. Carroll re-analyzed the data from scores of factor analytic studies from the preceding 50 years, and integrated the results into a single hierarchical structure of human abilities. However, long before that volume was published, factor analysis had mutated into a number of related procedures used for completely different purposes.

5.2. *Analysis of covariance structures*

In a wide-ranging yet brief review of the history of structural equation modeling, Bentler (1986) attributes to Tucker (1955) the origin of the distinction between "exploratory" and "confirmatory" factor analytic studies. For most of its early history (setting aside the founding article by Spearman that was intended to test his hypothesis about g), factor analysis was an essentially exploratory procedure with an Achilles' heel of rotational indeterminacy.

Bock and Bargmann (1966), in an article entitled *Analysis of Covariance Structures*, described an ML estimation algorithm, and a goodness of fit statistic, for a model formally very much like a factor analysis model but used for a totally different purpose. In matrix notation, the Bock and Bargmann model was identical to the factor analysis model, but some of the matrix elements that are estimated in the factor analysis model were fixed *a priori* to particular values (0s and 1s). The result (in the particular special case described by Bock and Bargmann) was a model for learning, but more general applications soon became apparent.

Jöreskog (1969) described a general approach to confirmatory maximum likelihood factor analysis, in which any of the usual parameters of the model (the factor loadings and/or the unique variances) could be estimated or fixed at a priori values. A first interesting aspect of this advance is that the computer programs based on Jöreskog's algorithm could be used to fit a number of special case models, e.g., Bock and Bargmann's learning curve model, with suitable placement of a priori fixed values. A second aspect of this development was that, if a sufficient number of fixed values are in appropriately chosen locations in the traditional factor analysis model, the long-standing problem of rotational indeterminacy disappears. Thus it became possible to statistically test the goodness of fit of a hypothesized factor-analytic model (including a particular rotation of the axes). Factor analysis thereby graduated from a purely exploratory procedure to become a truly confirmatory, statistical model-fitting activity. Furthermore, the techniques of confirmatory factor analysis can be used for many of the original purposes of exploratory factor analysis if the analyst is willing to specify a priori some (perhaps modest amount) of the structure.

The development of algorithms for ML estimation for (essentially factor-analytic) covariance structure models also became a steppingstone on what then turned out to be a very short crossing to the creation of structural equation modeling.

5.3. *Structural equation modeling*

Structural equation modeling (SEM) represents a marriage of the factor analysis model to multiple regression, in which regression brought Sewall Wright's (1921) path analysis as her dowry. Given the development of systems for the analysis of covariance structures, or confirmatory factor analysis, it was (in hindsight) a short step to the idea that the (confirmatory) factor analysis model could be used as the "measurement model" for a set of multiple indicators of a latent variable, and those latent variables could simultaneously be involved in multiple regression relations in a "structural model" (Jöreskog and van Thillo, 1973). The structural model describes a hypothesized network of directional and nondirectional relationships among latent variables. Jöreskog (1973), Wiley (1973), and Keesling (1972) developed the idea of SEM nearly simultaneously; Bentler and Weeks (1980) subsequently described an alternative model with essentially the same features.

Software to fit structural equation models is extremely general, and may be used to estimate the parameters, and test the fit, of a wide variety of more specific models. The models described in the previous section for analysis of covariance structures, or confirmatory factor analysis, are subsets of the linear structural equations model, as is the original exploratory factor analysis model. In addition, regression models may be fitted so as to account for measurement error in the observed variables (on either side of the equation) in the sense that the response and predictor variables in the regression(s) are the error-free unobserved, latent variables underlying the observed variables. This feature enabled statistically optimal estimation of the path coefficients that had been desired for path analysis since Wright (1921) first described the procedure. Fitting path models has become one of the most widespread uses of SEM.

The SEM model also subsumed the higher-order factor model, in which correlations among the factor scores of a classical (Thurstone) factor model are, in turn, subjected to

factor analysis. Thurstone (1947) had described the concept, and Schmid and Leiman (1957) had developed a multi-stage estimation procedure that involved repeated applications of then-current methods of factor extraction. However, because the factor model is, itself, a regression model, albeit one involving latent variables, the higher-order factor model may be expressed as a regression model involving phantom latent variables (that are not directly related to any observed variables). So expressed, the parameters of the higher-order factor model may be simultaneously estimated using SEM software, allowing the goodness of fit of the model to be evaluated.

In addition, by borrowing the strategy employed by Bock and Bargmann (1966) that placed fixed values in model matrices originally intended to hold estimated parameters, SEM software may be used to fit other models as well. Meredith and Tisak (1990) described the use of the SEM model to fit latent growth models, in which the covariance structure among repeated measurements is analyzed to yield an estimate of the functional form of the unobserved (error-free) growth process. (The analysis of growth involves extension of the SEM model to include the means of the repeated measurements; that extension of SEM has, in turn, had other uses as well.) A similar use of the same strategy permits SEM software to be used to fit variance component models (Jöreskog and Sörbom, 1989). The addition of mechanisms to fit SEMs simultaneously in multiple populations (Jöreskog and Sörbom, 1983) permits SEM software to be used to examine issues of factorial invariance of the kind earlier considered by Meredith (1964).

For the past 30 years, SEM has arguably been the area of most rapid development in quantitative psychology. The generality of the model has permitted its use in countless substantive studies in psychology, as well as in sociology, economics, and other fields. This widespread application has, in turn, motivated extensive methodological work, on topics such as parameter estimation with less restrictive distributional assumptions about the data, and (many) goodness of fit statistics. The story about SEM is not, by any means, written. Still, at this point it can be stated confidently that SEM, offspring of factor analysis that it is, will remain one of the most prominent contributions of quantitative psychology both to research in psychology and to the methods of applied statistics in general.

6. Psychological statistics

From the early part of the 20th century, psychometricians have made contributions to pedagogy in statistics, and to applied statistics more generally. The developments in psychological scaling, test theory, and factor analysis discussed in the preceding sections had as prerequisites the knowledge of fundamental concepts of statistics now taught in the first statistics courses that are universally required in undergraduate and graduate programs in psychology. Psychometricians produced volumes that were among the first texts for such courses; one of the earliest was E.L. Thorndike's (1913) volume *An Introduction to the Theory of Mental and Social Measurements* (2nd edition), which integrated what we would now call a first course in statistics with basic ideas of test theory. (The first edition of Thorndike's book appeared in 1904, but copies of that are now

relatively rare.) In the 1920s, Truman Kelley's (1923) *Statistical Methods*, Arthur Otis's (1926) *Statistical Method in Educational Measurement*, and Henry Garrett's (1926) *Statistics in Psychology and Education* appeared. Later texts by Guilford (1942), Edwards (1946) and McNemar (1949) anticipated in their content coverage nearly all subsequent textbooks for the first course in statistics in psychology and education. The everyday work of psychometricians in academic psychology departments has to this day included instruction in applied statistics courses, and countless additional undergraduate and graduate textbooks have been produced.

A driving force for change in instruction and practice in applied statistics over the course of the 20th century was the development of computing machinery. In the early part of that century, publications in statistics (and the practice of statistics) were dominated by the development and implementation of computational strategies, and algorithms for hand computation and the use of nomographs (graphical devices to assist computation) (Appelbaum, 1986). But computing machinery changed that, and psychometricians were again in the forefront. Ledyard Tucker's (1940) description of his modification of the "International Business Machines Corporation scoring machine" to perform matrix multiplication was likely the first publication in *Psychometrika* to describe a non-desk calculator method for statistical computation (Appelbaum, 1986). The modified "scoring machine" used marked multiple-choice answer sheets that had by then become common for large-scale test administration as the input medium, as well as the plug board; it was used in L.L. Thurstone's Psychometric Laboratories at the University of Chicago and at The University of North Carolina.

Psychometricians have been especially interested in providing computational methods for statistical procedures that are broadly useful in psychological research. Appelbaum (1986) notes that perhaps the longest-running topic in the history of *Psychometrika* involved the computation of the tetrachoric correlation that forms the basis of many approaches to item analysis in test theory. Research in psychology yields multivariate data in many experimental and observational contexts; in work that began in the 1960s at the Psychometric Laboratory at the University of North Carolina and continued at the University of Chicago in the 1970s, R. Darrell Bock developed matrix-based algorithms for a number of multivariate linear statistical procedures that were described in the influential text *Multivariate Statistical Methods in Behavioral Research* (Bock, 1975). Algorithms from that volume were implemented in several widely distributed computer programs, and the combination of the text for graduate courses and the computer applications led to widespread use of multivariate methods in psychological research.

Another class of multivariate data analysis techniques on which Bock (1960) also did some work is that of "optimal scaling." Bock's student, Shizuhiko Nishisato, is one of many psychometricians who have developed over the past few decades the theory and computational techniques of these procedures; overlapping sets of these techniques are also known as "dual scaling," "Hayashi's theory of quantification," "*analyse factorielle des correspondences*" ("correspondence analysis"), and "homogeneity analysis" (Nishisato, 1996). These procedures share with psychological scaling and test theory the capacity to quantify categorical data of kinds that appear in research in psychology: categorical item responses, ranks, and paired comparisons (to list a few of the more

common). While these procedures are related to the original (metric) multidimensional scaling, their scope has so increased that their heritage is largely obscured.

Closely related work by the Gifi (1990) group at the University of Leiden extends computationally intensive analysis of multivariate data using nonlinear transformations of nominal, ordered, or continuous data to provide nonlinear analogues of such multivariate techniques as regression and canonical correlation, and for cluster analysis. This work and more recent extensions (see Meulman, 2003; Heiser, 2004) often makes use of computer graphics to render visualizations of summaries of high-dimensional multivariate behavioral data that would otherwise be challenging to comprehend.

In work that some might say extends quantitative psychology's interest in multivariate statistics to spaces of infinite dimensionality, Ramsay and Silverman's (1997) *Functional Data Analysis* makes the unit of statistical analysis curves (such as growth curves) or surfaces (such as might describe some response as a function of more than one variable). In this form of analysis as well, the results are often provided graphically.

Indeed, the "results" of many forms of data analysis that have been the special province of psychometrics (psychological scaling, MDS, and various forms of optimal scaling) are essentially spatial representations of some data that are presented graphically, as curves or surfaces. As statistical graphics are at the interface between psychology and statistics (that is, between human perception and the data), quantitative psychologists have often been involved in research and development of statistical graphics even quite aside from the data analytic techniques that may underlie the graphs (Wainer and Thissen, 1993). Leland Wilkinson's (1999) *The Grammar of Graphics* integrates statistical graphics, computer science, and perception in a description of graphics that is essentially psycholinguistic in character (although its language is graphical). Tufte's (1983) volume, *The Visual Display of Quantitative Information*, and Wainer's (2005) *Graphic Discovery: A Trout in the Milk and Other Visual Adventures*, bring ideas about graphical data analysis to a broader audience.

Quantitative psychologists often have collaborated with statisticians to apply statistics to psychological research (e.g., Abelson, Bock, Bush, Green, Hedges, Jones, and Luce in combinations with Bargmann, Mosteller, Olkin, Tukey, and others). And quantitative psychology has also been concerned throughout its history with the topics of experimental design and data analysis. Roger Kirk's (1995) third edition of his classic *Experimental Design: Procedures for the Behavioral Sciences* stands as a monument to the training of graduate students in psychology for nearly four decades. Maxwell and Delaney's (1990) *Designing Experiments and Analyzing Data: A Model Comparison Perspective* integrates modern ideas of research design and data analysis. And in the past two decades, two new topics have become associated with research on experimental design and subsequent data analysis: On the design side, before data are collected, power analysis (Cohen, 1988) has become routine; that process is sufficiently challenging for the complex designs used in behavioral research that it has become its own subspecialty. The same is true for research synthesis (Cooper and Hedges, 1994), also sometimes referred to as meta-analysis (Glass, 1981; Hedges and Olkin, 1985), which seeks to statistically combine the results of collections of experimental studies to describe their common results and the features of variation they exhibit.

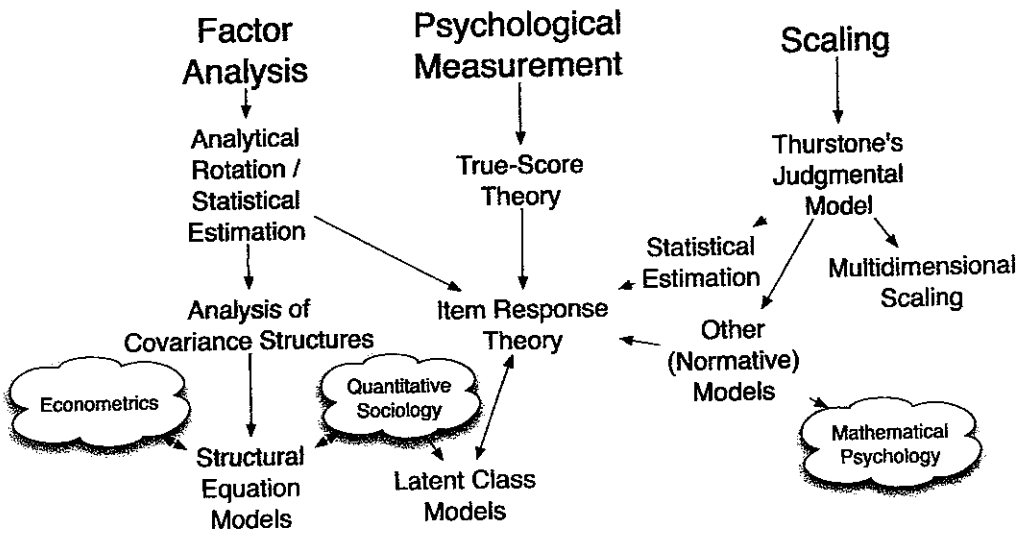

Fig. 2. A graphical description of some of the relations among the traditions of factor analysis, psychological measurement, and scaling and their extensions and developments.

Countless other topics of statistical research could be mentioned that have either captured the interest of psychometricians or been introduced in the literature of quantitative psychology. Suffice it to say that psychometrics has not been merely a consumer of the tools of statistics, but rather it has been one of the contributing fields to applied statistics throughout its long history and is likely to continue to be so.

7. Conclusion

Psychometrics, or quantitative psychology, is the disciplinary home of a set of statistical models and methods that have been developed primarily to summarize, describe, and draw inferences from empirical data collected in psychological research. The set of models and methods that are psychometric may be divided into three major classes. The first of these is psychological scaling, a set of techniques for the assignment of quantitative values to objects or events using data obtained from human judgment. A second class involves a large set of methods and procedures derived from the basic idea of factor analysis – to explain the observed covariation among a set of variables by appeal to the variation on underlying latent (unobserved) random variables that may explain the covariation. A third class of models has produced test theory; while the traditional algebraic true score theory is not a statistical model *per se*, the more recent item response theory is a class of statistical models that combines basic components from psychological scaling and latent explanatory variables from the factor analytic tradition. Figure 2 shows a graphical description of some of the relations among the traditions of factor analysis, psychological measurement, and scaling and their extensions and developments.

In addition, psychometrics has been the source of many other contributions to applied statistics, motivated by new data analytic challenges that have arisen with each new generation of scientific psychology.

We have provided some illustrations of the work of quantitative psychologists in the preceding four sections of this chapter. However, these illustrations have been selective; no attempt has been made to be exhaustive. Indeed, for every reference in this chapter, a dozen or more equally important publications are likely to have been omitted. We have sought to define psychometrics by providing some illustrations of what psychometricians do and have done, not to provide a comprehensive list of either topics or scholars.

While the intellectual history of psychometrics may be traced to roots two centuries in the past, and its existence as an identifiable discipline has been clear for seven decades, quantitative psychology remains a work in progress. Psychometric research has always drawn its motivation from contemporary challenges of psychological and behavioral research, and its tools from statistics. Because both the challenges of psychological research and the available statistical framework change with time, psychometrics continues to mature as a discipline. Subsequent chapters in this volume provide more detailed descriptions of many of the techniques that have been mentioned in this chapter, along with other novel developments.

References

Albert, J.H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics* **17**, 251–269.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. American Psychological Association, Washington, DC.

Anastasi, A., Urbina, S. (1996). *Psychological Testing*, 7th ed. Prentice-Hall, New York.

Angoff, W.H. (1971). Scales, norms and equivalent scores. In: Thorndike, R.L. (Ed.), *Educational Measurement*. 2nd ed. American Council on Education, Washington, DC, pp. 508–600. Reprinted as Angoff, W.H. (1984). Scales, Norms and Equivalent Scores. Educational Testing Service, Princeton, NJ.

Appelbaum, M.I. (1986). Statistics, data analysis, and *Psychometrika*: Major developments. *Psychometrika* **51**, 53–56.

Béguin, A.A., Glas, C.A.W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika* **66**, 541–561.

Bentler, P. (1986). Structural equation modeling and *Psychometrika*: A historical perspective on growth and achievements. *Psychometrika* **51**, 35–51.

Bentler, P.M., Weeks, D.G. (1980). Linear structural equations with latent variables. *Psychometrika* **45**, 289–307.

Binet, A., Simon, T. (1905). Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *Annee Psychologique* **11**, 191–244.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In: Lord, F.M., Novick, M.R. (Eds.), *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA, pp. 395–479.

Bloxom, B. (1974). An alternative method of fitting a model of individual differences in multidimensional scaling. *Psychometrika* **39**, 365–367.

Bock, R.D. (1960). Methods and applications of optimal scaling. The University of North Carolina Psychometric Laboratory Research Memorandum No. 25.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika* **37**, 29–51.

Bock, R.D. (1975). *Multivariate Statistical Methods in Behavioral Research*. McGraw-Hill, New York.

Bock, R.D. (1997a). The nominal categories model. In: van der Linden, W., Hambleton, R.K. (Eds.), *Handbook of Modern Item Response Theory*. Springer, New York, pp. 33–50.

Bock, R.D. (1997b). A brief history of item response theory. *Educational Measurement: Issues and Practice* **16**, 21–33.

Bock, R.D., Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. *Psychometrika* **46**, 443–459.

Bock, R.D., Bargmann, R. (1966). Analysis of covariance structures. *Psychometrika* **46**, 443–449.

Bock, R.D., Jones, L.V. (1968). *The Measurement and Prediction of Judgment and Choice*. Holden-Day, San Francisco, CA.

Bock, R.D., Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika* **35**, 179–197.

Bock, R.D., Moustaki, I. (this volume), Item response theory in a general framework. Chapter 15 in this volume.

Boring, G.W. (1929). *A History of Experimental Psychology*. Appelton-Century, New York.

Bradley, R.A., Terry, M.E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* **39**, 324–345.

Bradlow, E.T., Wainer, H., Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika* **64**, 153–168.

Brennan, R.L. (2001). *Generalizability Theory*. Springer, New York.

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

Brown, W., Thomson, G.H. (1921). *The Essentials of Mental Measurement*. Cambridge University Press, Cambridge.

Browne, M.W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research* **36**, 111–150.

Burt, C. (1922). *Mental and Scholastic Tests*. P.S. King, London.

Carroll, J.B. (1953). An analytical solution for approximating simple structure in factor analysis. *Psychometrika* **18**, 23–38.

Carroll, J.B. (1993). *Human Cognitive Abilities: A Survey of Factor Analytic Studies*. Cambridge University Press, Cambridge.

Carroll, J.D., Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart–Young" decomposition. *Psychometrika* **35**, 283–319.

Cattell, J.M. (1890). Mental tests and measurements. *Mind* **15**, 373–380.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ.

Cooper, H., Hedges, L. (1994). *The Handbook of Research Synthesis*. Russell Sage Foundation. New York.

Cremin, L.A. (1961). *The Transformation of the School: Progressivism in American Education, 1876–1957*. Alfred A. Knopf, New York.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* **16**, 297–334.

Cronbach, L.J. (1990). *Essentials of Psychological Testing*. Harper & Row, New York.

Cronbach, L.J., Gleser, G.C., Nanda, H., Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, New York.

Cronbach, L.J., Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin* **52**, 281–302.

DuBois P.H. (1947). The classification program. Army Air Forces Aviation Psychology Program Research Reports, Report No. 2. U.S. Government Printing Office, Washington, DC.

Edwards, A.L. (1946). *Statistical Analysis for Students in Psychology and Education*. Rinehart, New York.

Feldt, L.S., Brennan, R.L. (1989). Reliability. In: Linn, R.L. (Ed.), *Educational Measurement*. 3rd ed. American Council on Education/Macmillan, New York, pp. 105–146.

Feuer, M.J., Holland, P.W., Green, B.F., Bertenthal, M.W., Hemphill, F.C. (Eds.) (1999). *Uncommon Measures: Equivalence and Linkage Among Educational Tests*. National Academy Press, Washington, DC.

Fiske, D.W., Hanfmann, E., Mackinnon, D.W., Miller, J.G., Murray, H.A. (1993). *Selection of Personnel for Clandestine Operations: Assessment of Men*. Aegean Park Press, Walnut Creek, CA.

Flanagan, J.C. (1948). The aviation psychology program in the army air forces. Army Air Forces Aviation Psychology Program Research Reports, Report No. 1. U.S. Government Printing Office, Washington, DC.

Flanagan, J.C. (1951). Units, scores, and norms. In: Lindquist, E.F. (Ed.), *Educational Measurement*. American Council on Education, Washington, DC, pp. 695–763.

Fox, J.P., Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* **66**, 269–286.

Fox, J.P., Glas, C.A.W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika* **68**, 169–191.

Garrett, H.E. (1926). *Statistics in Psychology and Education*. Longmans, Green and Co., New York.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Wiley, New York.

Glass, G.V. (1981). *Meta-Analysis in Social Research*. Sage, Beverly Hills, CA.

Goddard, H.H. (1908). The Binet and Simon tests of intellectual capacity. *Training School Bulletin* **5** (10), 3–9.

Green, B.F. Jr. (1954). Attitude measurement. In: Lindzey, G. (Ed.), *Handbook of Social Psychology*, vol. I. Addison-Wesley, Cambridge, MA, pp. 335–369.

Guilford, J.P. (1936). *Psychometric Methods*. McGraw-Hill, New York.

Guilford, J.P. (1942). *Fundamental Statistics in Psychology and Education*. McGraw-Hill, New York.

Guilford, J.P., Lacey, J.L. (Eds.) (1947). Printed classification tests. Army Air Forces Aviation Psychology Program Research Reports, Report No. 5. U.S. Government Printing Office, Washington, DC.

Gulliksen, H.O. (1950). Theory of Mental Tests. Wiley, New York. Reprinted in 1987 by Lawrence Erlbaum Associates, Hillsdale, NJ.

Gulliksen, H. (1961). Measurement of learning and mental abilities. *Psychometrika* **26**, 93–107.

Handler, L. (2001). Assessment of men: Personality assessment goes to war by the Office of Strategic Services assessment staff. *Journal of Personality Assessment* **76**, 558–578.

Harmon, H.H. (1976). *Modern Factor Analysis*, 3rd ed., revised. University of Chicago Press, Chicago.

Hedges, L.V., Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, Orlando, FL.

Heiser, W.J. (2004). Geometric representation of association between categories. *Psychometrika* **69**, 513–545.

Hilgard, E.R. (1987). *Psychology in America: A historical survey*. Harcourt Brace Jovanovich, Orlando, FL.

Holland, P.W., Dorans, N.J. (in press). Linking and equating test scores. In: Brennan, R.L. (Ed.), *Educational Measurement*, 4th ed. In press.

Holland, P.W., Rubin, D.B. (1982). *Test Equating*. Academic Press, New York.

Horan, C.B. (1969). Multidimensional scaling: Combining observations when individuals have different perceptual structures. *Psychometrika* **34**, 139–165.

Horst, P. (1941). A non-graphical method for transforming an arbitrary factor matrix into a simple structure factor matrix. *Psychometrika* **6**, 79–99.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika* **6**, 153–160.

Jennrich, R.I., Robinson, S.M. (1969). A Newton–Raphson algorithm for maximum likelihood factor analysis. *Psychometrika* **34**, 111–123.

Jennrich, R.I., Sampson, P.F. (1966). Rotation for simple loadings. *Psychometrika* **31**, 313–323.

Jones, L.V., Olkin, I. (2004). *The Nation's Report Card: Evolution and Perspectives*. Phi Delta Kappa International, Bloomington, IN.

Jöreskog, K.G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika* **32**, 443–482.

Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* **34**, 183–202.

Jöreskog, K.G. (1973). A general method for estimating a linear structural equation system. In: Goldberger, A.S., Duncan, O.D. (Eds.), *Structural Equation Models in the Social Sciences*. Academic Press, New York, pp. 85–112.

Jöreskog, K.G., Sörbom, D. (1983). *LISREL User's Guide*. International Educational Services, Chicago.

Jöreskog, K.G., Sörbom, D. (1989). *LISREL 7 User's Reference Guide*. Scientific Software, Chicago.

Jöreskog, K.G., van Thillo, M. (1973). LISREL – A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables. Research Report 73-5. Uppsala University, Department of Statistics, Uppsala.

Kaiser, H.F. (1958). The varimax criterion for analytical rotation in factor analysis. *Educational and Psychological Measurement* **23**, 770–773.

Keesling, J.W. (1972). Maximum likelihood approaches to causal analysis. Unpublished Doctoral Dissertation. University of Chicago.

Kelley, T.L. (1914). Comparable measures. *Journal of Educational Psychology* **5**, 589–595.

Kelley, T.L. (1923). *Statistical Methods*. Macmillan, New York.

Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Science*, 3rd ed. Brooks/Cole, Pacific Grove, CA.

Kolen, M.J., Brennan, R.L. (2004). *Test Equating, Linking, and Scaling: Methods and Practices*, 2nd ed. Springer-Verlag, New York.

Koretz, D.M., Bertenthal, M.W., Green, B.F. (Eds.) (1999). *Embedding Questions: The Pursuit of a Common Measure in Uncommon Tests*. National Academy Press, Washington, DC.

Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**, 1–27.

Kruskal, J.B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **29**, 115–129.

Kuder, G.F., Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika* **2**, 151–160.

Lawley, D.N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh* **62-A** (Part I), 74–82.

Lawley, D.N., Maxwell, A.E. (1963). *Factor Analysis as a Statistical Method*. Butterworth, London.

Lawley, D.N., Maxwell, A.E. (1971). *Factor Analysis as a Statistical Method*, 2nd ed. Butterworth, London.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. In: Stouffer, S.A., Guttman, L., Suchman, E.A., Lazarsfeld, P.F., Star, S.A., Clausen, J.A. (Eds.), *Measurement and Prediction*. Wiley, New York, pp. 362–412.

Leuchtenburg, W.E. (1958). *The Perils of Prosperity, 1914–32*. The University of Chicago Press, Chicago.

Lewis, C. (1986). Test theory and *Psychometrika*: The past twenty-five years. *Psychometrika* **51**, 11–22.

Lingoes, J.C. (1965). An IBM program for Guttman–Lingoes smallest space analysis. *Behavioral Science* **10**, 183–184.

Lord, F.M. (1952). A Theory of Test Scores. *Psychometric Monographs*, Whole No. 7.

Lord, F.M. (1953a). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement* **13**, 517–548.

Lord, F.M. (1953b). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika* **18**, 57–76.

Lord, F.M., Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA.

Luce, R.D. (1959). *Individual Choice Behavior*. Wiley, New York.

Luce, R.D., Tukey, J.W. (1964). Simultaneous conjoint measurement. *Journal of Mathematical Psychology* **1**, 1–27.

Masters, G.N., Wright, B.D. (1984). The essential process in a family of measurement models. *Psychometrika* **49**, 529–544.

Maxwell, S.E., Delaney, H.D. (1990). *Designing Experiments and Analyzing Data: A model Comparison Perspective*. Wadsworth, Belmont, CA.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika* **29**, 177–185.

Meredith, W., Tisak, J. (1990). Latent curve analysis. *Psychometrika* **55**, 107–122.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist* **35**, 1012–1027.

Meulman, J.J. (2003). Prediction and classification in nonlinear data analysis: Something old, something new, something borrowed, something blue. *Psychometrika* **68**, 493–517.

McNemar, Q. (1949). *Psychological Statistics*. Wiley, New York.

Mulaik, S.A. (1986). Factor analysis and *Psychometrika*: Major developments. *Psychometrika* **51**, 23–33.

Nishisato, S. (1996). Gleaning in the field of dual scaling. *Psychometrika* **61**, 559–599.

Office of Strategic Services Assessment Staff (1948). *Assessment of Men*. Rinehart, New York.

Otis, A.S. (1916). The reliability of spelling scales, including a 'deviation formula' for correlation. *School and Society* **4**, 96–99.

Otis, A.S. (1918). An absolute point scale for the group measurement of intelligence. *Journal of Educational Psychology* **9**, 239–261, 333–348.

Otis, A.S. (1926). *Statistical Method in Educational Measurement*. World Book Company, Yonkers-on-Hudson, NY.

Patz, R.J., Junker, B.W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics* **24**, 146–178.

Patz, R.J., Junker, B.W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics* **24**, 342–366.

Petersen, N.S., Kolen, M.J., Hoover, H.D. (1989). Scaling, norming and equating. In: Linn, R.L. (Ed.), *Educational Measurement*. 3rd ed. Macmillan, New York, pp. 221–262.

Ramsay, J.O., Silverman, B.W. (1997). *Functional Data Analysis*. Springer, New York.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Denmarks Paedagogiske Institut, Copenhagen. Republished in 1980 by the University of Chicago Press.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, CA, pp. 321–333.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology* **19**, 49–57.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In: Blegvad, M. (Ed.), *The Danish Yearbook of Philosophy*. Munksgaard, Copenhagen.

Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores. *Psychometric Monographs*, Whole No. 17.

Samejima, F. (1997). Graded response model. In: van der Linden, W., Hambleton, R.K. (Eds.), *Handbook of Modern Item Response Theory*. Springer, New York, pp. 85–100.

Schmid, J., Leiman, J.M. (1957). The development of hierarchical factor solutions. *Psychometrika* **22**, 53–61.

Shepard, R.N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika* **27**, 125–140.

Shepard, R.N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika* **27**, 219–246.

Sokal, M.M. (1982). James McKeen Cattell and the future of anthropometric mental testing. In: Woodward, W.R., Ash, M.G. (Eds.), *The Problematic Science: Psychology in Nineteenth-Century Thought*. Praeger Publishers, New York, pp. 322–345.

Spearman, C. (1904). "General intelligence", objectively determined and measured. *American Journal of Psychology* **15**, 201–293.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology* **3**, 271–295.

Stevens, S.S. (1951). Mathematics, measurement, and psychophysics. In: Stevens, S.S. (Ed.), *Handbook of Experimental Psychology*. Wiley, New York, pp. 1–49.

Symonds, P.M. (1929). Choice of items for a test on the basis of difficulty. *Journal of Educational Psychology* **20**, 481–493.

Terman, L.M. (1916). *The Measurement of Intelligence*. Houghton Mifflin, Boston.

Thissen, D., Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika* **51**, 567–577.

Thorndike, E.L. (1913). *An Introduction to the Theory of Mental and Social Measurements*. Teachers College, Columbia University, New York.

Thurstone, L.L. (1924). *Psychological Examination for High School Graduates and College Freshmen*. American Council on Education, Washington, DC.

Thurstone, L.L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology* **16**, 433–449.

Thurstone, L.L. (1927). The law of comparative judgment. *Psychological Review* **34**, 278–286.

Thurstone, L.L. (1931a). Multiple factor analysis. *Psychological Review* **38**, 406–427.

Thurstone, L.L. (1931b). *The Reliability and Validity of Tests*. Edwards Brothers, Ann Arbor, MI.

Thurstone, L.L. (1935). *Vectors of the Mind*. University of Chicago Press, Chicago.

Thurstone, L.L. (1937). Psychology as a quantitative rational science. *Science* **85**, 227–232.

Thurstone, L.L. (1938). Primary Mental Abilities. *Psychometric Monographs*, Whole No. 1. University of Chicago Press, Chicago.

Thurstone, L.L. (1947). *Multiple Factor Analysis*. University of Chicago Press, Chicago.

Torgerson, W.S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika* **17**, 401–419.

Torgerson, W.S. (1958). *Theory and Methods of Scaling*. Wiley, New York.

Torgerson, W.S. (1986). Scaling and *Psychometrika*: Spatial and alternative representations of similarity data. *Psychometrika* **51**, 57–63.

Tucker, L.R. (1940). A matrix multiplier. *Psychometrika* **5**, 289–294.

Tucker, L.R. (1946). Maximum validity of a test with equivalent items. *Psychometrika* **11**, 1–13.

Tucker, L.R. (1955). The objective definition of simple structure in linear factor analysis. *Psychometrika* **20**, 209–225.

Tucker, L.R. (1972). Relations between multidimensional scaling and three-mode factor analysis. *Psychometrika* **37**, 3–27.

Tucker, L.R., Messick, S. (1963). An individual differences model for multidimensional scaling. *Psychometrika* **28**, 333–367.

Tufte, E.R. (1983). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT.

van der Linden, W.J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement* **22**, 195–211.

van der Linden, W.J., Glas, C.A.W. (2000). *Computerized Adaptive Testing: Theory and Practice*. Kluwer Academic Publishers, Dordrecht.

von Davier, A., Holland, P.W., Thayer, D. (2003). *The Kernel Method of Test Equating*. Springer, New York.

Wainer, H. (2005). *Graphic Discovery: A Trout in the Milk and Other Visual Adventures*. Princeton University Press, Princeton, NJ.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B.F., Mislevy, R.M., Steinberg, L., Thissen, D. (2000). *Computerized Adaptive Testing: A Primer*, 2nd ed. Lawrence Erlbaum Associates, Mahwah, NJ.

Wainer, H., Thissen, D. (1993). Graphical data analysis. In: Keren, G., Lewis, C. (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Statistical Issues*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 391–457.

Wang, X., Bradlow, E.T., Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement* **26**, 109–128.

Wiley, D.E. (1973). The identification problem for structural equation models with unmeasured variables. In: Goldberger, A.S., Duncan, O.D. (Eds.), *Structural Equation Models in the Social Sciences*. Academic Press, New York, pp. 69–83.

Wilkinson, L. (1999). *The Grammar of Graphics*. Springer, New York.

Wissler, C. (1901). The correlation of mental and physical tests. *Psychological Review Monograph Supplements* **3** (6).

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research* **20**, 557–585.

Yerkes, R.M. (1921). Psychological Testing in the US Army. *Memoirs of the National Academy of Sciences*, vol. XV.

Young, F.W., Torgerson, W.S. (1967). TORSCA, a Fortran IV program for nonmetric multidimensional scaling. *Behavioral Science* **13**, 343–344.

Young, F.W., Hamer, R.M. (1987). *Multidimensional Scaling: History, Theory, and Applications*. Lawrence Erlbaum Associates, Hillsdale, NJ.