2

The History of Educational Measurement

Brian E. Clauser
NBME (Retired)

Jerome C. Clauser
American Board of Internal Medicine

Amanda L. Clauser
NBME

Clauser, B. E., Clauser, J. C., & Clauser, A. L. (2025). The history of educational measurement. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 33–113). Oxford University Press.
DOI: 10.1093/oso/9780197654965.003.0002

The history of educational measurement is not a simple history of evolving theory leading to the current state of the art. It is, in fact, multiple separate but related histories. There is a history of test theory and a related history defined by the key figures in a range of fields. Some of these figures, such as Charles Spearman, made critical contributions to test theory; others, such as Walter Bingham and Lewis Terman, influenced testing practice but are likely to go unmentioned in a discussion of test theory. Gustav Fechner studied psychophysics, Alfred Binet studied child development (and mesmerism), and Spearman explored the nature of intelligence; together, they provide a significant part of the foundation on which educational measurement has developed. Other important contributors, such as Truman Kelley, appear to have been singularly focused on the development of methodology.

In addition to these parallel histories built around theory and important contributors to the field, there is a history of evolving technology. Computers have become so widespread in test administration that someone new to the field might overlook the impact that the optical scanner had on testing practice during the second half of the 20th century. In addition to the obvious impact that technology has had on the scoring and administration of tests, computational power has profoundly impacted psychometric theory and practice. In the days when *computers* were individuals using slide rules or mechanical calculators, even a correlation coefficient might have required hours—or days—of work. Pearson (1907), Spearman (1904a), and Kuder and Richardson (1937) all put substantial effort into developing computationally simple approaches to approximating correlation (or reliability) coefficients. The increase in computational power since the 1930s has made such approximations seem quaint and has brought with it applications of Bayesian statistics, hierarchical models, and machine learning that have made educational measurement into something that would be almost unrecognizable to early practitioners.

There has also been an evolution in the application of testing. Spearman was interested in correlations across tests within populations as a means of understanding the nature of intelligence (Spearman, 1904a). Much early large-scale testing—in the military and industry—focused on personnel selection. In this context, tests evaluated individual differences. This same perspective accompanied the introduction of large-scale testing programs in schools, including not only achievement tests, which remain common, but also group IQ tests. Admissions tests follow a similar model; more recently, considerable emphasis has been given to national and international assessments for which individual scores may not even be calculated.

Finally, there is also a history related to the social trends and pressures that have provided the context in which educational measurement has evolved. The Darwinian revolution of the second half of the 19th century led to the eugenics movement, which had widespread support until the Second World War. Much of educational measurement in the early 1900s became inextricably tied to that movement. Similarly, it does not seem like unjustified speculation to suggest that the emphasis that Cronbach (1971, 1989b) and Messick (1975, 1989) placed on fairness and consequences as part of test validity

can be accounted for (at least in part) by the social context created by the civil rights movement and the war on poverty. These social pressures have also included an ebb and flow of the presence of the antitesting movement.

In considering the history of educational measurement, a similar challenge is raised by the question of where to begin. The use of examinations to select individuals for government positions in China began more than 3,000 years ago. In 1115 BCE, candidates were tested in the "six arts": music, archery, horsemanship, writing, arithmetic, and rites and ceremonies of public and private life. During the Han dynasty—approximately 2,000 years ago—written examinations were introduced in the "five studies": civil law, military affairs, agriculture, revenue, and geography. After the 7th century, national civil service tests were introduced for use across the empire. These tests emphasized the ability to remember and interpret the Confucian classics. They used both written and oral formats (Zhang & Luo, 2020). Modeled at least in part on these Chinese tests, in 1833 a testing program was introduced for selecting individuals for the British Indian civil service. Testing at universities as part of the process of granting degrees has also been an established practice for centuries (DuBois, 1970).

The breadth and depth of this history makes it clear that a comprehensive history is well beyond the scope of this chapter. Instead, we have identified specific parts of that history which we believe are relevant for understanding the current state of educational measurement. In selecting the parts we chose to stress, we considered several competing priorities. Most people find origin stories interesting. We hope that the events and individuals we discuss create a compelling story that will motivate some readers to explore this history in more detail. In this regard, we hope our chapter will provide a basis for basic historical literacy related to the field of educational measurement. In addition, we have attempted to provide context for understanding some of the major developments in both the theory and the practice of educational measurement. We are more likely to use measurement theories appropriately if we know the context in which they were developed. The reliability coefficient has become a ubiquitous part of test theory and practice, but our reliance on this coefficient may well have as much to do with the history of the field as it does with the usefulness of the coefficient (Cronbach & Shavelson, 2004). Additionally, viewing measurement and testing in the context of an early history that is closely tied to intelligence testing and eugenics is likely to help us understand the skepticism of and resistance to testing that have been present since the 1920s.

It is likely too much to hope that understanding the role that the misuse of measurement has played in supporting injustice will ensure that measurement is not misused in the future, but such an understanding is likely to help us remember that we need to ask not only what we *can* do with our technology, but also what we *should* (and should not) do.

With these considerations in mind, we have divided this chapter into eight sections, each with a different focus. The first section provides a discussion of some of the early foundations of educational measurement, including the contributions of Charles Spearman, Gustav Fechner, and Alfred Binet. To understand how Spearman's contributions

came into being, we begin with the intellectual culture of the middle of the 19th century that led Francis Galton to study heredity and how Galton's use of correlation in that study provided the statistical tool that Spearman would subsequently apply to test scores. Fechner is important because his work on psychophysics represents what is likely the first scientific effort to measure psychological phenomena. Similarly, Binet introduced not only the intelligence test, but also, along with Theodore Simon, the idea that plots of performance against an independent variable (in this case, age rather than proficiency) can be used in test development.

The second section examines the rise of intelligence testing in the United States. We begin with the work of Henry Goddard and Lewis Terman in the first quarter of the 20th century. Their early work positioned them to be leaders in the development of the army testing program during World War I. That program provided a significant impetus for the rise of testing in the United States. The rise of intelligence testing in turn provided impetus for the eugenics movement in the United States.

The third section provides an overview of the impact that U.S. government programs have had on shaping the practice of educational measurement. This includes the related military research carried out during and after World War I. It also includes legislation that has funded education and both mandated and defined the scope of educational assessment.

The fourth through sixth sections trace the development of three critical aspects of test theory: classical test theory, scaling, and item response theory (IRT). The fourth section follows the development of classical test theory, picking up where the first section left off. This includes the work of Kelley, G. Frederic Kuder, and Marion Richardson, as well as Lee Cronbach's work on coefficient alpha, concluding with a discussion of the development of generalizability theory.

The fifth section builds on the discussion that was begun in the first section, which introduced the development of scaling through the contributions of Fechner and Binet. This section continues the story, following it through to the work of Edward Thorndike and Louis Thurstone, who together laid a kind of theoretical foundation for IRT. The sixth section discusses the independent but parallel development of IRT and Rasch measurement.

The seventh section again shifts from measurement theory to practice, examining the history of the use of large-scale tests both for selection and for accountability. We begin with the tests that Horace Mann introduced to evaluate the performance of the Boston school system in the first half of the 19th century. We then show how this type of test evolved to produce assessments like the College Board examinations and truly large-scale assessments, including the SAT (formerly Scholastic Aptitude Test) and the National Assessment of Educational Progress (NAEP).

The eighth and final section in this chapter shows how the five editions of *Educational Measurement* both document and have been a part of the history of educational measurement.

EARLY FOUNDATIONS OF EDUCATIONAL MEASUREMENT

As we noted in the introduction, educational measurement has a long history. Written standardized tests for selection of government officials were administered in China 3,000 years ago. Written tests have been used at universities for hundreds of years—the competitive mathematics test administered at Cambridge has been in place since the 18th century. Tests for the British Indian civil service began in 1833. These early tests for selection and graduation laid a foundation for the developments that followed. They used standardized conditions of administration and emphasized fairness and accuracy in scoring, but as DuBois (1970) noted, testing during this period lacked the theoretical framework and statistical techniques that have become the hallmark of educational measurement. Those conceptualizations had their roots in England, Germany, and France in work that spanned the period from 1860 through 1910.

England can appropriately be viewed as the birthplace of correlational psychology, and much of what we refer to as classical test theory has its beginning in the work of Charles Spearman (1863–1945), the best known of the early proponents of correlational psychology. Germany was the birthplace of experimental psychology. Gustav Fechner developed psychophysics procedures at the University of Leipzig and, in the process, provided what was likely the first scientific measure of psychological phenomena. Finally, in France, Alfred Binet developed an intelligence test with the practical goal of identifying children in need of special schooling.

Developments in England

In Harold Gulliksen's 1950 text, he summarized the history of classical test theory in a single sentence: "Nearly all the basic formulas that are particularly useful in test theory are found in Spearman's early papers" (p. 1). In a very real sense, Spearman created the foundation for what we know as mathematical test theory, so we begin this history by describing the historical context that brought Spearman's formulas into being.

The chain of events that led to Spearman's work began nearly half a century earlier and halfway around the globe. In 1858, Alfred Russel Wallace (1823–1913) was working in the Malay Archipelago. Wallace supported himself by collecting natural history specimens to be sold in England; he was also collecting evidence in an effort to answer the question of how species come into being. That year, he wrote a paper titled *On the Tendency of Varieties to Depart Indefinitely From the Original Type* (Wallace, 1858). That paper provided an elegant summary of Charles Darwin's—unpublished—theory of natural selection. Wallace mailed his manuscript to Darwin (1809–1882) with the request that if he thought it was of merit, he should forward it to the eminent geologist Charles Lyell. At the point at which Darwin received the letter, he had spent much of the previous 2 decades working (largely in secret) on the same theory and collecting material for a planned multivolume work on the topic.

The events that followed the arrival of Wallace's letter are well known. Darwin shared Wallace's paper with Lyell and Joseph Hooker, two men who were both preeminent members of the British scientific community and Darwin's closest friends. They arranged to have Wallace's paper published with two short pieces by Darwin (Brooks, 1984). All of this has importance to the history of test theory because it made Darwin realize that he could no longer delay publication of a more complete statement of his theory. In just over a year, the first edition of *On the Origin of Species* was published (Darwin, 1859). That book profoundly impacted the history of science; it also profoundly impacted Darwin's first cousin, Francis Galton.

Francis Galton

Francis Galton (1822–1911) was the quintessential Victorian polymath. He studied meteorology, created some of the first weather maps, and discovered the phenomenon of the anticyclone. He investigated the use of fingerprints for identification and wrote three books that were instrumental in the adoption of fingerprint technology (Galton, 1892, 1893, 1895). He explored a part of Africa previously unknown to Europeans (Galton, 1853). He wrote extensively on eugenics (Galton, 1909); in fact, he coined the term. But of all the seemingly limitless areas that Galton explored, his greatest contributions came from his study of heredity motivated by his reading of Darwin's *Origin*.

Galton had begun his intellectual life by studying medicine at King's College London. He had pursued medicine at the urging of his father; he also studied mathematics at Cambridge. After his father's death, Galton used his inherited wealth to mount an expedition to Africa. When he returned, he published a book on his travels (Galton, 1853) and was awarded a medal by the Royal Geographical Society. But as Stephen Stigler (1986) wrote, "Darwin's theories opened an intellectual continent" (p. 267) for Galton to explore far greater than Africa; Galton spent decades mapping that intellectual continent. This exploration of heredity led him to identify the phenomenon of regression to the mean and develop the powerful analytic tool represented by *correlation*.

Galton began his work on inheritance by documenting the extent to which exceptional talent was shared across generations in the same family. In *Hereditary Genius* (1869), he recorded instance after instance in which notable individuals (in the arts, science, and jurisprudence) have notable relatives in the same field. He used this evidence to argue that intellectual and personality characteristics can be inherited in the same way that physical characteristics are passed from parent to child. From a contemporary perspective, the book appears simplistic; Galton too quickly discounted the impact of environmental factors and personal advantage as explanations for his results. Within the context of the times, the work must have been more impressive. Darwin appears to have overlooked these alternative explanations when he wrote to Galton (Darwin, 1869, first paragraph):

> I have only read about 50 pages of your Book (to the Judges) but I must exhale myself, else something will go wrong in my inside. I do not think I ever in all my life read anything more interesting & original. And how well & clearly you put every point!

Galton was an astute observer; he also carried out organized experiments to understand inheritance. In 1875, he experimented on pea plants. (Gregor Mendel's work had been published 9 years earlier in an obscure journal [Mendel, 1866], but remained unknown to the broader European scientific community.) Galton described the experiment in his autobiography (Galton, 1908).

> The following question had been much in my mind. How is it possible for a population to remain alike in its features, as a whole, during many successive generations, if the average produce of each couple resemble their parents? . . . I was very desirous of ascertaining the facts of the case. After much consideration and many inquiries, I determined . . . on experimenting with sweet peas, which were suggested to me both by Sir Joseph Hooker and by Mr. Darwin. . . . I procured a large number of seeds from the same bin, and selected seven weights.
> I persuaded friends living in various parts of the country, each to plant a set for me. . . . The result clearly proved *Regression*; the mean Filial deviation was only one-third that of the parental one, and the experiments all concurred. The formula that expresses the descent from one generation of a people to the next, showed that the generations would be identical if this kind of *Regression* was allowed for. (pp. 300–302)

Galton followed the sweet pea experiments with efforts to collect data on humans. He solicited help from friends to collect physical measurements on parents and children and on siblings. He also opened the anthropometric laboratory at the International Health Exhibition in London. This allowed him to collect a range of measures on individuals who visited the exhibition (Galton, 1884).

These data collection efforts allowed Galton to refine his understanding of regression, and in 1886 he contributed two papers on familial relations to the Royal Society (Galton, 1886a, 1886b). These focused his attention on tables representing the relation between deviations in measures such as stature of the adult child and the same measures for parents. Galton realized that a relation existed between the measures, but he was at a loss to quantify it. Again, he described the event in his autobiography: "At length, one morning, while waiting at a roadside station near Ramsgate for a train, and poring over the diagram in my notebook, it struck me that the lines of equal frequency ran in concentric ellipses" (Galton, 1908). He returned to London and visited the Royal Institution in search of information on conic sections. At the Royal Institution, a chance encounter with the physicist James Dewar led to the suggestion that Dewar's brother-in-law, the mathematician J. Hamilton Dickson, might be able to help. Dickson viewed it as a simple problem and his solution was presented as an appendix to Galton's resulting paper.

Two years later, Galton (1888) introduced the correlation coefficient. At that time, the term and general concept of correlation were well established. As Galton stated,

> "co-relation or correlation of structure" is a phrase much used in biology, and not least in that branch of it which deals with heredity . . . ; but I am not aware of any previous attempt to define it clearly, to trace its mode of action in detail, or to show how to measure its degree. (p. 135)

The 1888 paper built on his earlier work on regression. Again, using anthropological data, he showed how a coefficient could be produced by placing the measures of the correlated attributes on the same scale. This coefficient had the advantage that:

> (1) $y = rX$ for all values of y; (2) that r is the same, whichever of the two values is taken as the subject; (3) that r is always less than 1; (4) that r measures the closeness of co-relation. (p. 145)

Galton had succeeded not only in making an important contribution to the understanding of heredity, but also in providing the mathematical tools to explore a wide range of phenomena. Although we have found no references to indicate that Galton was aware of Spearman's use of the correlation coefficient, if he was aware of Spearman's work, he likely would not have been surprised. In the introductory chapter of *Natural Inheritance*, Galton (1889) commented that it would be worth the reader's time to understand the methodology he used in his study of heredity because of its wide application:

> It familiarizes us with the measurement of variability, and with the curious laws of chance that apply to a vast diversity of social subjects. This part of the inquiry may be said to run along a road on a high level, that affords wide views in unexpected directions, and from which easy descents may be made to totally different goals to those we have now to reach. (p. 3)

Galton's great contribution to statistical science—and the history of educational measurement—was the phenomenon of regression to the mean and the powerful analytic tool correlation. There is, of course, some controversy about attributing the discovery of correlation to Galton, but at a minimum, he deserves credit for pointing out the broad usefulness of the procedure in the social and biological sciences. His own words written in another context describe his role in the history of correlation well: "It is a most common experience that what one inventor knew to be original, and believed to be new, had been invented independently by others many times before, but had never become established" (Galton, 1889, p. 33).

In addition to making direct contributions to the development of various fields of science, Galton impacted scientific development through the influence he had on the next generation of thinkers. One individual strongly influenced by Galton was Francis Edgeworth; Edgeworth made important contributions to the development of the correlation coefficient and wrote what may be the first discussion of the impact of measurement error on test scores; we will discuss his work in detail in a subsequent section. Another important follower of Galton was Karl Pearson.

Karl Pearson

Galton's concept of correlation is critical to classical test theory, but the specifics of the coefficient Galton created have long since been replaced by the more mathematically tractable form presented by Karl Pearson. (Galton's approach was based on median and interquartile distance rather than mean and variance.) Although Pearson spent relatively little time studying psychological measurement, his development of the product moment correlation coefficient and his relationships with Galton and Spearman make him a critical part of this history nonetheless.

Like Galton, Pearson (1857–1936) studied mathematics at Cambridge. Unlike Galton, who suffered a breakdown attempting to earn an honors degree, Pearson excelled in those studies, ranking third "wrangler" in the competitive examination taken by all Cambridge undergraduates studying math(s).[1] Pearson was an incredibly productive researcher and author: The annotated bibliography of his works contains 648 entries (Morant & Welch, 1939). Those hundreds of publications included important contributions to statistics; he is responsible for the chi squared test (Pearson, 1900) and for the product moment correlation coefficient (Pearson, 1896). Pearson presented the correlation coefficient in the following form,

$$R = S(xy)/(n\sigma_1\sigma_2) \quad (1)$$

crediting others with developing this equation; he used the bulk of his 1896 paper to argue for the advantages of this formula.[2]

Pearson devoted much of his professional life to advancing Galton's ideas. With Galton's support, he founded and edited the journal *Biometrika* to advance the statistical study of evolutionary biology. After Galton's death, Pearson held the Galton chair in eugenics, created with an endowment from Galton.

Pearson's admiration for Galton is clear in the following quotation from a speech Pearson delivered at a dinner in his own honor in 1934:

> "Road on a high level," "wide views in unexpected directions," "easy descents to totally different goals"—here was a field for an adventurous roamer! I felt like a buccaneer of Drake's days. . . . I interpreted that sentence of Galton to mean that there was a category broader than causation, namely correlation, of which causation was only the limit, and that his new conception of correlation brought psychology, anthropology, medicine and sociology in large parts into the field of mathematical treatment. It was Galton who first freed me from the prejudice that sound mathematics could only be applied to natural phenomena under a category of causation. Here for the first time was a possibility—I will not say a certainty of reaching knowledge—as valid as physical knowledge was then thought to be—in the field of living forms and above all in the field of human conduct. (Pearson, 1934, pp. 22–23)

This speech also provides a vivid reminder of another part of Galton's legacy that Pearson embraced:

> Buccaneer expeditions into many fields followed; fights took place on many seas, but whether we had right or wrong, whether we lost or won, we did produce some effect. The climax culminated in Galton's preaching eugenics, and his foundation of the Eugenics Professorship. Did I say "culmination"? No, that lies rather in the future, perhaps with Reichskanzler Hitler and his proposals to regenerate the German people. In Germany a vast experiment is in hand, and some of you may live to see its results. (Pearson, 1934, p. 23)[3]

Correlational Psychology in 1900

Eight years after Pearson published his paper on the product moment correlation coefficient, Spearman began publishing the papers that Gulliksen referred to as containing the formulas that are particularly useful in test theory. The content of those papers warrants special attention, but before examining that content, it is appropriate to consider the state of correlational psychology immediately before the publication of Spearman's papers. Spearman's methodological contributions to educational measurement were enormous, but he was not the first to apply correlational methods to psychology.

The first of these earlier papers we consider was written by William Bagley. Bagley (1901) tested schoolchildren using a range of what might be called physical and mental tests. In general, he found no relationship (or inverse relationships) between these measures. He specifically noted, "There seems to be little direct relation between mental ability as represented by reaction times, and mental ability as represented by class standings" (Bagley, 1901, p. 205). Although Bagley's paper has the title *On the Correlation of Mental and Motor Ability in School Children*, his approach to "correlation" was simplistic even in the context of the times. He examined the relationships between his measures by dividing the first measure into quintiles and averaging the values within each quintile. He then averaged the corresponding values of the second variable for the students in those same quintiles. The five pairs of numbers were then examined directly or plotted to determine whether any relationship existed.

Two additional papers from *Psychological Review* also provide examples of the prior application of correlation in psychology. Clark Wissler's Columbia University dissertation was published in 1901. S. Lovie and Lovie (2010) suggested that this was likely the first use of the product moment correlation coefficient published in a psychology journal. Wissler's paper was an early example of the now-proud tradition of using undergraduates in psychology research. Starting with the hypothesis that different measures of intelligence would be correlated, he reported on numerous measures of physical and mental proficiency, including reaction time, accuracy in marking out As in text, speed in marking out As in text, accuracy in drawing a line, length and breadth of head, and class standing in numerous subjects. Contrary to Wissler's hypothesis, the strength of these relationships was modest except for that between class standing in different subjects. Wissler recognized that "failure to

correlate" may be "due to want of precision in the tests" (Wissler, 1901; p. 29) but gave no serious consideration to how reliability might be calculated or used to adjust the correlation.

The second of the two papers (Aikens et al., 1902) also presents relations between measures of proficiency. The research shows a remarkable lack of statistical sophistication. The "correlations" that are presented have no clear relationship to the Pearson product moment approach. Instead, they used what might be described as a makeshift approximation. Again, the results show generally modest relations between the measured proficiencies, and again, the paper lacks any consideration of how to evaluate the impact of systematic or random effects that might impact the correlation of interest.

Charles Spearman

As a young man, Spearman had an interest in Indian philosophy and joined the army with the hope of being stationed in India. In his 1930 autobiographical chapter, Spearman described the nearly 15 years he spent in the army as wasted, although it did give him the opportunity to complete a 2-year "staff college." Not long after completing these studies, he resigned his commission and moved to Leipzig to study psychology with Wilhelm Wundt. His studies were interrupted when he was recalled to the army at the outbreak of the Boer War, but eventually he returned to Leipzig, completed his doctorate, and subsequently became a professor at University College London (P. Lovie & Lovie, 1996; S. Lovie & Lovie, 2010; Spearman, 1930).

Spearman's work focused on understanding the nature of intelligence. Although his most enduring contribution to science may be the methodological tools he developed for correlational psychology and test theory, these tools were developed to support empirical evaluation of the nature of intelligence and specifically to explore the theory that there is a general factor that underlies all measures of intelligence. Differences in reliability across tests impacted the observed correlations between test scores. Much of Spearman's methodological work was focused on accounting for this effect.

We began this discussion of the developments that took place in England with Gulliksen's reference to Spearman's early papers (Gulliksen, 1950). The first of these papers was "The Proof and Measurement of Association Between Two Things" (Spearman, 1904b). This was followed later in the same year by "'General Intelligence' Objectively Determined and Measured" (Spearman, 1904a). Both were published while Spearman was still working on his dissertation in Leipzig. Interestingly, these papers were unrelated to his dissertation work, which focused on spatial localization, a topic more in line with Wundt's program of research (Jensen, 1998; Tinker, 1932). The first paper introduced variations on the rank-order correlation coefficient and argued for the advantages of correlation based on ranks rather than measures. It also presented Yule's (1897) formula for partial correlation, which Spearman stated he derived using

an independent approach. Most importantly, the paper introduced the formula for estimating true score correlation (p. 98)[4]:

$$r_{pq} = \frac{r_{p'q'}}{\sqrt{r_{p'p'} \cdot r_{q'q'}}}. \tag{2}$$

In this equation, r_{pq} represents the true score correlation between variables p and q, $r_{p'q'}$ is the observed correlation between p and q, and the denominator on the right side of the equation represents the square root of the product of the reliabilities for p' and q'.

In discussing the importance of correction that this formula produced, Spearman quoted a paper in which Pearson had concluded that "the mental characteristics in man are inherited in precisely the same manner as the physical" (Spearman, 1904b, pp. 97–98). Spearman pointed out that if the observed correlations between relatives are similar (which Pearson reported), the actual relationship must be substantially stronger for the psychological characteristics because they are measured less precisely. This criticism started a feud between the two University College London professors that continued for nearly a quarter of a century.

Pearson's immediate response to Spearman's criticism was to republish the lecture Spearman cited, adding an addendum responding to the criticism (Pearson, 1904). The addendum attacked Spearman's correction for attenuation because it was capable of producing results that exceed unity and called on Spearman to provide algebraic proofs of the formulas. Spearman responded to Pearson's criticism in 1907 with a "Demonstration of Formulae for True Measurement of Correlation." That same year, Pearson published a paper on methods for estimating correlation and again attacked Spearman's work. In Pearson's 39-page report (Pearson, 1907), Spearman's name appears more than 20 times. This creates the impression that the real motivation behind the publication may have had more to do with attacking Spearman than with advancing the practice of correlational science.

The last of Spearman's papers to make a major contribution to test theory, "Correlation Calculated From Faulty Data," was published in 1910. It again responded to Pearson's criticisms and, more importantly, presented the Spearman–Brown formula,

$$r_{x[p],x[p]} = \frac{p \cdot r_{x[q],x[q]}}{q + (p-q) r_{x[q],x[q]}}. \tag{3}$$

In this formulation presented by Spearman (1910, p. 281), $r_{x[q],x[q]}$ represents the reliability of a test with $2qi$ items (with i being any number), and $r_{x[p],x[p]}$ represents the reliability when the test comprises $2pi$ items.[5] This formula represents a seminal contribution to test theory. When the test length is doubled, the formula provides a basis for estimating split-half reliability. As we will see, this framework led to conceptualizing reliability in terms of interitem correlation and ultimately led to coefficient alpha.

Immediately following Spearman's paper in the *British Journal of Psychology* was a paper by William Brown (1881–1952; Brown, 1910a), which presents an alternative proof for the same formula. Because of this, the names Spearman and Brown have

been linked as though they were coauthors; that is an ironic misconception. The work presented in Brown's paper comes from his then–recently completed dissertation. The dissertation was also published by the author (Brown, 1910b) and in modified form as *The Essentials of Mental Measurement* (Brown, 1911). Much of the dissertation—and the subsequent book—was an attack on Spearman. Both publications so carefully documented Pearson's criticisms of Spearman that even without Brown's acknowledgment—"Professor Pearson has very kindly read the entire thesis in proof, and made several useful suggestions" (Brown, 1910b, p. 1)—it would be clear that Pearson and not Spearman had influenced Brown's work.

Spearman's 1910 paper is a response to Pearson's criticisms, and he clearly hoped to minimize their differences, concluding, "On the whole, if we eliminate these misapprehensions and oversights, there seems to be no serious difference of opinion on all these points between Pearson and myself" (p. 288). If this was intended as an olive branch, it was one that Pearson did not accept; Pearson continued his attacks for years to come (e.g., Pearson & Moul, 1927).

That said, there is one area in which Pearson and Spearman seem to have been in agreement: They both fully supported the eugenics movement. The following quote concisely captures Spearman's views: "One can conceive the establishment of a minimum index (of general intelligence) to qualify for parliamentary vote, and above all, for the right to have offspring" (Hart & Spearman, 1912, p. 79).

Spearman's Immediate Impact on Measurement

Despite Pearson's criticism, Spearman's ideas became widely accepted, although it is difficult to make a clear statement about how fast that acceptance came about. Edward Thorndike, Wilfrid Lay, and P. R. Dean published a paper in 1909 calling the conclusions of Spearman (1904a) into question, but that same paper applied Spearman's correction for unreliability without comment. The authors appeared to accept Spearman's methodology while rejecting the idea that the available evidence supported the existence of "general intelligence."

Two years later, A. R. Abelson—also at University College London—published an extensive study on the development of a battery of tests. In this paper, Abelson used Spearman's methodology and provided what may be the first example of the use of reliability coefficients to determine the number of items to include on a test form (Abelson, 1911).

By the middle of the 1920s, numerous researchers had published empirical studies evaluating the accuracy of the Spearman–Brown formula. Using a variety of tests, the researchers compared the predicted reliabilities to actual results produced by changing the test length: Holzinger and Clayton (1925) did so using the Otis advanced mental test, Ruch et al. (1926) used spelling words, and B. D. Wood (1926) used achievement tests. The results left little doubt about the usefulness of the formula. Remmers et al. (1927) extended these findings by applying the Spearman–Brown formula to rating

scales and showing that the formula predicted the change in reliability as a function of the number of raters completing the scale.

As we will discuss in a subsequent section, the formulations provided by Spearman did much to shape thinking about reliability for the next 50 years. Numerous researchers made substantial contributions, but the general approach based on correlation coefficients and a conceptualization of reliability as split-half reliability remained intact.[6]

Psychophysics and the University of Leipzig

It would be difficult to write a history of educational measurement and not mention the University of Leipzig. The fact that Charles Spearman received his doctorate from Leipzig would be enough to ensure inclusion, but the university made a foundational contribution to what would become educational measurement well before Spearman arrived. Psychophysics can reasonably be viewed as the starting point for the empirical study of psychology. Gustav Fechner's (1801–1887) 10-year project that culminated in his *Elements of Psychophysics* (1860/1966) may reasonably represent the first effort to measure the psychological. In the process, he laid the foundations for psychological scaling. Stigler (1999) also credited Fechner's work as being the reason that psychology was the first of the social sciences to incorporate statistical theory.

Fechner was born in 1801 in Lower Lusatia, part of the Holy Roman Empire. He moved to Leipzig at the age of 16 to study medicine and stayed for the rest of his life. For the history of educational measurement, and more broadly psychology, the critical events took place in 1850 when Fechner apparently had a kind of epiphany. As Stigler (1986) stated, the stories of that epiphany are "colorful and dramatic" (p. 243). The dramatic aside, Fechner's interest not only in physics but also in philosophy and psychology led him to think about the relationship between physical energy and psychological energy; his insight was that if it is possible to measure physical energy, it must also be possible to quantify the corresponding psychological energy. Fechner took inspiration from the work of Ernst Weber (1795–1878), who was also at the University of Leipzig. Weber had done earlier work on discriminating between physical stimuli and had concluded that the magnitude of a just noticeable difference was a fixed proportion of the base stimulus (Boring, 1950). The study of just noticeable differences can take many forms. For example, a subject might be asked to lift two weights—which outwardly appear identical—and report on whether they are of equal weight or if one is heavier than the other. Obviously, if the difference in the two weights is small enough, it will generally not be perceived. As the difference increases, the probability that it will be perceived increases. Weber's law (sometimes known as the Weber–Fechner law) posits that the ratio of the difference between the two weights (ΔI) and the base weight (I) is a constant (k):

$$\frac{\Delta I}{I} = k. \tag{4}$$

Put another way, as the base weight of the objects increases, the difference in weight between the objects must increase proportionally to have the same probability of being

noticed. As we will discuss in more detail in the section on scaling, Fechner modified this relationship by basing it on the logarithm of the stimulus.

For the bulk of Fechner's work in this area, he experimented with discriminating between weights. He used apparently identical containers, one of which represented the base weight and the other of which would have an additional weight placed inside. For a given pair of weights, the relevant metric was the number of correct responses divided by the number of trials. The resulting Weber-Fechner law was a central result reported in the *Elements of Psychophysics* (Fechner, 1860/1966), but the real contributions of Fechner's work are methodological. Over time, both the accuracy and the generalizability of the law have been called into question, but the general approach Fechner used provided a model for subsequent work in psychophysics, and more importantly, Fechner introduced research design and statistical analysis into the practice of psychology.

Fechner recognized that a wide variety of factors could impact his results: which hand was used, which weight was presented first, the subject's level of fatigue. To account for these sorts of effects, he used balanced factorial designs that allowed him to not only account for the effect across repetitions of the experiment, but also directly estimate the effect of different conditions. For example, he might have a subject lift the same two weights with both the right and the left hands. Although Fechner did not use random sampling, Stigler described the *Elements of Psychophysics* (Fechner, 1860/1966) as the most comprehensive treatment of experimental design to appear before R. A. Fisher's (1935) *Design of Experiments* was published 75 years later.

In imagining Fechner's (1860/1966) experiments, it may be easy for the reader to envision a laboratory in which undergraduates queued up to participate in replications in which they lifted weights. This would be far from the reality. Much of Fechner's work was carried out with Fechner as the only subject—a subject who knew which container contained which weight.

After Fechner retired from lecturing, Wilhelm Wundt came to the University of Leipzig. Wundt continued the psychophysical program of research and established the Institute for Experimental Psychology. This would become the world's first graduate program in psychology.

Alfred Binet

In the early 1900s, when Spearman was publishing the papers that provided the foundation for test theory, Alfred Binet (1857–1911) was publishing a series of papers that provided a foundation for the practice of intelligence testing. Although Binet worked as a psychologist for decades, these papers and the scale they describe are the reason his name is so widely recognized.

Binet studied law; he received his license in 1878, but was unhappy with his career choice and began reading psychology at the Bibliotheque nationale, where he became familiar with Galton's work. In 1882, he met Jean Martin Charcot and Charles Féré and began working in Charcot's laboratory observing and experimenting with the effects of hypnosis on hysteria. In 1887, Binet published a monograph with Féré entitled *Le mag-*

netisme animal. During that year, he also became a student at his father-in-law's laboratory of embryology at the College de France. At the laboratory, he spent several hours per week practicing dissection. The results of this work provided the basis for Binet's doctoral dissertation, *A Contribution to the Study of the Subintestinal Nervous System of Insects* (Wolf, 1973).

During this time, Binet acted as an outspoken supporter of Charcot's ideas about hypnosis. As the debate over these ideas became more heated, the relatively young Binet found himself in an unenviable and untenable position and eventually severed his relationship with Charcot. In search of a new institutional affiliation, Binet requested a staff position at the Laboratory of Physiology and Psychology at the Sorbonne. In 1894, he became director of the laboratory and in collaboration with the previous director founded the first French psychology journal, *L'année psychologique* (Wolf, 1973).

During the latter part of his career, Binet had several less senior collaborators; the most noteworthy of these was Theodore Simon. When Simon approached Binet, he had completed medical school studies and needed a thesis to finalize his degree. At the time, Simon was an intern at the colony for children and adolescents with intellectual disabilities at Perray-Vaucluse. Rather than accept Simon into the laboratory outright, he challenged him to return to the colony and take a series of measurements of each of the 223 boys in the institution and report descriptive statistics for each measure by age group. Several months later, Simon returned, with his analyses completed, and was accepted by Binet; the measurements formed the basis of the thesis for his medical degree.

At the same time that Simon became Binet's assistant, Binet was asked to become part of the Societe libre pour l'etude psychologique de l'enfant. The society was subsequently asked to form a commission on the education of children with intellectual disabilities. Part of the commission's charge was to develop a means of identifying children who would benefit from being educated outside the regular classroom. As Binet and Simon described it (1916),

> In October 1904 the Minister of Public Instruction named a commission which was charged with the study of measures to be taken for insuring the benefits of instruction to defective children. . . . They decided that no child suspected of retardation should be eliminated from the ordinary school and admitted to a special class, without being subjected to a pedagogical and medical examination from which it could be certified that because of the state of his intelligence, he was unable to profit, in an average measure, from the instruction given in the ordinary schools. (p. 9)

They go on to say that such decisions about placement

> have at all times too much the nature of the arbitrary, of caprice, of indifference. Such a condition is quite unfortunate because the interests of the child demand a more careful method. To be a member of a special class can never be a mark of distinction, and such as do not merit it, must be spared the record. (pp. 9–10)

As a leader of the society, Binet was appointed to a commission and spent much of the remainder of his career (and his life) working on this problem.

Between 1905 and 1911, Binet and Simon published a series of papers in *L'année psychologique* that outlined the problem and recorded the effort they made to develop an instrument to identify children with intellectual disabilities.[7] They began by reviewing the state of the art for such diagnosis, arguing that at the time diagnosis was typically carried out by physicians who based their conclusions on physical symptoms. Binet and Simon asserted that causes such as birth trauma or hydrocephalus may result in disability, but they do not explain the extent of the intellectual deficit. They cited a number of diagnostic frameworks with categories such as *hydrocephalic idiocy, epileptic idiocy*, and *traumatic idiocy.* They then argued that even when previous researchers had provided diagnostic criteria based on observable manifestations of intelligence, the criteria had been too vague to be useful. For example, one system stated that "in profound idiocy '*the attention is fugitive,*' while in imbecility, '*the attention is fleeting*'" (Binet & Simon, 1916; p. 23).

In their next paper, also in 1905 (included in Binet & Simon, 1916), Binet and Simon laid out a synthetic approach including psychological, pedagogical, and medical methods. They began their description of the psychological method by admitting the limitations on measurement when it is applied to intelligence: "The scale properly speaking does not permit the measure of intelligence, because intellectual qualities are not super-posable, and therefore cannot be measured as linear surfaces" (Binet & Simon, 1916; p. 40).

They noted that previous researchers had built their approaches on theory. By contrast, Binet and Simon made practical decisions in constructing their scale. They tried many tests and retained those that proved useful. They designed their individual tests to be administered quickly to avoid fatigue or loss of interest on the part of the child. Finally, they described what we would now refer to as standardized conditions for the tests: "The examination should take place in a quiet room"; "the child should be called in without other children"; when the child is introduced to the experimenter, "he should be reassured by the presence of someone he knows"; and so on (Binet & Simon, 1916; p. 44).

Binet and Simon then presented procedures for carrying out 30 tests. At the simple end of the continuum, the experimenter was instructed to move a lit match across the subject's field of vision and record whether he or she followed it visually. They assessed "recognition of food" by alternately presenting a piece of chocolate and a similarly shaped piece of white wood. The most sophisticated test required the subject to define abstract terms: for example, "What difference is there between weariness and sadness?"

Binet and Simon followed this description of their procedures with a 90-page report on applications of their new method. Binet's contribution to educational measurement may be viewed as more of a contribution to practice than to theory. That said, the presentation of the original Binet–Simon scale included methodological and conceptual breakthroughs that have been enormously important. The first of these is empirical eval-

uation of tests (items) for inclusion in the scale. Binet and Simon tried out numerous tests and retained those that discriminated between children of different ages. Although they did not have sophisticated mathematical models to guide them, their approach to evaluating test material was a precursor to the item characteristic curves that would become an essential part of IRT.[8]

The second conceptual contribution that Binet and Simon made was the idea of a norm-referenced interpretation of the scores. Although the samples that were used for the development of their original scale were small—often in the range of 8 to 15 children for a specific age group—they attempted to use children who would be described as average for their age. Admittedly, this seems quaint when compared to the approaches to norming examinations that were used later in the century, but conceptually Binet and Simon provided the starting point for this practice.

The remainder of the work on the Binet–Simon scale was published in 1908 and 1911 (see Binet & Simon, 1916). Both papers describe revisions. In the present context the details are unimportant, but there are two issues that remain salient in the context of intelligence testing. The first is the recognition that these tests are sensitive to education; they are not tests of innate intelligence. One motivation for revision was to remove (sub)tests that were particularly sensitive to education. The second issue is the extent to which scores are influenced by "social conditions." In the 1911 paper (included in Binet & Simon, 1916), they discussed the work of other researchers who applied their methods and found that many of the children performed well above their age level. They hypothesized that the differences resulted from the fact that the second set of children were from more affluent conditions. Clearly, concern about bias in intelligence testing has a long history.

One final contribution from Binet is an early conceptualization of what we now think of as adaptive testing. Binet recognized that there was little value in administering a test that was very difficult or very easy for the child. He instructed the administrator to move quickly through tasks that were easily mastered by the child or skip them altogether if other evidence from interacting with the child made it clear that the tests would not be useful. Clearly, this practical approach that focused on maintaining the child's attention is a far cry from the statistically sophisticated approaches that are currently available for adaptive testing, but it shows that an understanding of the value of matching item difficulty to test-taker proficiency was present from the early part of the 20th century.

Taken together, the work that went on over a century ago in England, Germany, and France did much to provide the theoretical foundation of educational measurement. In the next section, we describe how that foundation led to the widespread use of intelligence tests in the United States and how that use has impacted and continues to impact the field of educational measurement.

THE RISE OF INTELLIGENCE TESTING IN AMERICA

Clearly, the development of educational measurement in the 20th century was tightly linked to intelligence testing. Spearman's main interest in test scores was as a means

of understanding human intelligence, and Binet's motivation was the assessment of intelligence in schoolchildren. In addition to these more obvious connections, one of the earliest efforts to systematically quantify intelligence had its methodological roots in Leipzig.

James McKeen Cattell

In 1882, James McKeen Cattell (1860–1944) won a fellowship to study philosophy at Johns Hopkins. At Hopkins, he worked in G. Stanley Hall's psychology laboratory, where he measured the time in milliseconds it took subjects to recognize letters of the Latin alphabet (Sokal, 1990). After a year at Hopkins, Cattell moved to Leipzig to work on a doctorate at Wundt's laboratory, again studying reaction time. When he left Leipzig, he moved to England to study medicine. In England he met Galton, who was actively working on collection and analysis of anthropometric data (Galton, 1884). Both the reaction time work and anthropometric measurement influenced Cattell's contribution to the history of intelligence testing. In 1891, Cattell moved to Columbia University and established a psychology laboratory; 3 years later, he began testing each student entering Columbia College. The testing program went on into the beginning of the new century. Ultimately, some of these data were analyzed by Clark Wissler (1901) using product moment correlations. The results (which again showed few relations among scores from the different tests) put an end to Cattell's program of research and substantially deterred the use of psychophysical measurements in subsequent intelligence testing (Sokal, 1990). These results ended Cattell's testing program but did nothing to undermine the perceived need for measures of intelligence. As we discussed in the previous section, in 1905 Binet and Simon began publishing work to fill that need, and it was not long before Henry Goddard started to apply Binet and Simon's methods in the United States.

Henry Goddard

Like Cattell, Henry Goddard (1866–1957) studied with G. Stanley Hall, earning his doctorate under Hall in 1899. In 1906, Edward Johnstone, superintendent of the New Jersey School for Feeble-Minded Boys and Girls in Vineland, offered to build Goddard a laboratory if he would come to Vineland to conduct research. Goddard spent the next 2 years working with the diverse population at Vineland and came to conclusions similar to those described by Binet and Simon in their 1905 paper (see Binet & Simon, 1916). Both in France and in the United States the main responsibility for such children fell to physicians, and in both countries the diagnostic classifications and criteria were all but useless.

In 1908, Goddard set out on a tour of Europe in search of more useful approaches, and while in Belgium he came across the intelligence tests developed by Binet and Simon. Goddard's experience at Vineland put him in a position to see the value of Binet's work, and when he returned to New Jersey he translated the tests and began to use them. The resulting scale agreed with the impressions of the Vineland staff

who had direct experience with the children, and Goddard rapidly became a convert. By 1910, Goddard had used the Binet–Simon method extensively. That year, at Goddard's urging, the Classification Committee of the American Association for the Study of the Feeble-Minded adopted the method. This was a turning point for intelligence testing because it represented formal acceptance of the approach within the medical community (Zenderland, 1990, 1998).

Although Goddard's immediate challenge at Vineland had to do with decisions around educational placement for children, his interest in *feeblemindedness* broadened quickly. In 1911, he used the Binet–Simon test in public schools. By 1913, he had tried out the tests at Ellis Island (Zenderland, 1998). At that point, he had already published his eugenically motivated book, *The Kallikak Family* (Goddard, 1912). Kallikak was a pseudonym for a family that Goddard studied; at the time, Deborah Kallikak lived at Vineland. The book was a kind of mirror image of Galton's *Hereditary Genius*; instead of tracing recurrent instances of genius within the same family, Goddard traced instances of mental deficiency within a single family. Interestingly, although the book highlights the cost to society of such families, Goddard also advocated for the positive results that would be produced by what he viewed as appropriate education.[9]

Goddard was not alone in his interest in the Binet–Simon scale. In 1911, Lewis Terman published an account of his application of the scale with 400 children. And before Terman published the first version of the Stanford–Binet scale there were at least four other English translations and revisions (Hilgard, 1989). In addition to having continued in use through multiple revisions, Terman's version of the test popularized the use of the intelligence quotient (IQ).[10] This work by Goddard and Terman positioned them both to make contributions when the entrance of the United States into World War I led to a new application for intelligence testing.

Robert Yerkes and Mental Testing in the Army

There can be little doubt that the large-scale use of group intelligence tests during the First World War did much to raise the profile of psychology as a profession and testing as an accepted part of American life. Several authors have called into question the extent to which the test results were used to make important decisions about recruits (e.g., Gould, 1996; Minton, 1990), but there is no question that the army testing program demonstrated the potential for large-scale testing. Although numerous individuals shaped the project, it was led by Robert Yerkes (1876–1956). When Woodrow Wilson went before Congress to ask for a declaration of war, he also asked that universal conscription be instituted to provide the necessary manpower to wage war. Yerkes, the president of the American Psychological Association (APA), saw an opportunity. It is impossible to know to what extent he saw that opportunity as a chance for psychologists to contribute to the war effort and to what extent he may have been motivated by a desire to raise the

profile of psychology as a profession—and, with it, his own standing. Any combination of these motivations may have existed.

Weeks after the declaration of war and before the draft was approved, Yerkes called a meeting of the APA's executive council in Philadelphia. At the meeting, the council approved the formation of an APA Committee on Methods of Psychological Examining of Recruits. The committee met the next month at Vineland (Samelson, 1990; Yoakum & Yerkes, 1920). The group represented a who's who of intelligence testing, including Yerkes, Goddard, Walter Bingham, and Terman. At this time, intelligence tests were typically individually administered, and the initial plans were based on this model. The committee quickly realized that an alternative was needed and within weeks they had developed a prototype for a standardized, group-administered test that could be efficiently administered and scored.

Although the prototype for the army test was put together relatively quickly, it went through various stages of revision and evaluation. After revision, 10 tests remained. These tests were administered to approximately 5,000 men in the U.S. Army and National Guard, as well as samples from "institutions for the feeble-minded," officer training schools, and colleges and universities. The papers were then sent to Columbia University to be evaluated by a group led by E. L. Thorndike "to check their validity, reliability and significance" (Yoakum & Yerkes, 1920, p. 5).

The proposal for widespread testing of recruits with what was termed the Army Alpha was accepted and a school was established for training in military psychology; by the end of the war on November 11, 1918, approximately 120 officers, 300 enlisted men, and 500 clerks were involved in the examination effort. Yoakum and Yerkes (1920) stated (with impressive precision) that 1,726,966 men were examined.

The Army Alpha

The Army Alpha represented a pivotal point in the development of educational measurement in the United States. It represents the beginning of truly large-scale testing and widespread use of intelligence tests. It also provides results that fanned the flames of the eugenics movement and led to the antitesting movement.

As we noted, the committee working on the test quickly came to the realization that a group test was needed. This decision was critically important, but it was not unprecedented. In 1913, William Pyle published *The Examination of School Children*; this was likely the first group intelligence test. It was based on simple constructed response formats (Weinland, 1973).

A second important contribution that impacted the development of Army Alpha was made by Frederick Kelly. In 1915, Kelly published *The Kansas Silent Reading Test*. He was concerned about the unreliability of teachers' marks; he had written his dissertation on the topic. Kelly also hoped to reduce the time and effort required for administration and scoring of tests. Working at the State Normal School at Emporia, Kansas, and then as dean of education at the University of Kansas, he developed an item type that allowed for selection rather than construction of responses. The following two

items qualify as icons in the field of educational measurement in that they are two of the first multiple-choice questions ever published:

> Think of the thickness of the peelings of apples and oranges. Put a line around the name of the fruit having the thinner peeling.
>
> apples oranges
>
> Three words are given below. One of them has been left out of this sentence: I cannot _____ the girl who has the flag. Draw a line around the word which is needed in the above sentence.
>
> red see come

Although some parts of the Army Alpha required simple constructed responses—"How many men are 30 men and 7 men?" (Yoakum & Yerkes, 1920, p. 206)—most used the multiple-choice format.

A final contribution that impacted the development of the Army Alpha was made by Arthur Otis (1886–1964). In 1917, Otis was a graduate student at Stanford University, working with Terman. Otis had developed a group intelligence test and Terman brought a copy of the unpublished manuscript to Vineland (Yoakum & Yerkes, 1920). Yerkes's committee, working at Vineland, reviewed hundreds of published tests and ultimately modeled the Army Alpha on Otis's work. Otis interrupted his graduate studies to participate in the development and administration of Army Alpha. He published his version of the test in 1918 (Otis, 1918a, 1918b). After the war, Otis finished his degree and joined the World Book Company,[11] which published the Otis Group Intelligence Scale in 1920 and the Otis Self-Administering Tests of Mental Ability in 1922.

Terman, Peacetime Intelligence Testing, and Controversy

Whether the Army Alpha was used for important decisions and improved the efficiency of the war effort may be in question, but there is little doubt that the demonstration of large-scale intelligence testing provided support for the intelligence testing movement in the United States. Application of testing in the army took a back seat until the United States entered the Second World War, but group tests were rapidly adopted by school systems. By 1922, there were more than 40 group intelligence tests being published and in the 1922–1923 school year more than 3 million students were tested (Weinland, 1973). This number continued to rise and large-scale use of intelligence tests continued for decades.

In addition to publishing extensively and developing tests, in 1921 Terman began what is one of the longest-running longitudinal psychological studies in history.[12] Terman identified 1,000 "gifted" children from California with the intention of following them as they grew. The study was intended to provide information about the extent to which childhood IQ scores predicted later success in education and careers. The first results were reported in 1925 (Terman et al., 1925) and the sixth report on the gifted group in later maturity was published in 1995 (Holahan et al., 1995).[13]

As a, if not *the*, leading figure in intelligence testing at the time, Terman was soon at the center of the controversy that arose around testing. In 1922, Walter Lippmann—

author and cofounder of the *New Republic*—wrote a six-part attack on intelligence testing (Lippmann, 1922a, 1922b, 1922c, 1922d, 1922e, 1922f). Much of the content of the articles was specifically targeted at Terman. Lippmann (1922b) began with an effort to ridicule testing by focusing on the apparent absurdity of using the results of the Army Alpha test to conclude that the average American adult had the intelligence of a 14-year-old. Lippmann (1922c) then raised concerns about the potential of the tests to stigmatize children. He also expressed concern that rather than motivating remediation, low scores may provide justification for failing to provide the resources needed for education (Lippmann, 1922d). He moved on to argue against the conclusion that intelligence—as measured by these tests—is inherited and largely unchanged over time (Lippmann, 1922e). Lippmann concluded by suggesting that psychologists should abandon intelligence testing and "save themselves from the reproach of having opened up a new chance for quackery in a field where quacks breed like rabbits" (Lippmann, 1922f, p. 10).

Considering that they were written for the popular press, Lippmann's articles (1922a, 1922b, 1922c, 1922d, 1922e, 1922f) included a surprising amount of technical detail. He criticized the appropriateness of the norming group used for the Stanford–Binet (Lippmann, 1922a). He argued that the results of the Army Alpha were manipulated by the decisions made about the time limits for the individual tests and that it was impossible to conclude that the intelligence measured by the tests is inherited because so much learning occurs before the age of 4, when these tests were first administered (Lippmann, 1922b).

Terman's response to Lippmann began mockingly:

> After Mr. [William Jennings] Bryan had confounded the evolutionists, and [the flat earth advocate] Voliva the astronomers, it was only fitting that some equally fearless knight should stride forth in righteous wrath and annihilate that other group of pseudo-scientists known as "intelligence testers." Mr. Walter Lippmann, alone and unaided, has performed just this service. That it took six rambling articles to do the job is unimportant. It is done. The world is deeply in debt to Mr. Lippmann. So are the psychologists, if they only knew it, for henceforth they should know better than to waste their lives monkeying with those silly little "puzzles" or juggling IQ's and mental ages. (Terman, 1922, p. 116)

In responding to Terman, Lippmann moved from attacking testing to attacking Terman, with lines such as "Mr. Terman's logical abilities are so primitive that he finds this point impossible to grasp" (Lippmann, 1923, p. 146). Lippmann's views aside, intelligence testing and the view that intelligence was inherited continued to grow during the 1920s.

Intelligence Testing and Eugenics

Educational measurement had its foundation in the work of Galton and Spearman, both committed eugenicists. During the first 2 decades of the 20th century, the eugenics movement was becoming well established in both England and the United States. Although we are not aware of evidence to suggest that the Army Alpha was motivated by eugenics, after the war, when detailed accounts of the test were published (Yerkes, 1921; Yoakum & Yerkes, 1920), the results certainly fanned the flames of that move-

ment. Two types of results produced particular controversy. First, the very large sample of recruits allowed for examining differences across race and ethnicity. Substantial differences were apparent between Black and White recruits. Similar differences were also observed between recruits who had recently arrived in the United States from eastern and western Europe. These results could not have been more in line with the expectations of the eugenicists who believed that White western Europeans were of superior stock. Both Henry Goddard and Carl Brigham (1890–1943)—psychologists who had worked on the Army Alpha during the war—republished these results in the form of arguments for a eugenic solution. Brigham's (1923) *A Study of American Intelligence* was particularly strident.[14]

Goddard and Brigham were far from the only psychologists of this era who embraced eugenics. Truman Kelley was a member of the advisory council to the Eugenics Committee of the United States of America (I. Fisher, 1923). Psychologists testified before Congress in support of restrictive immigration laws (although there is little evidence that their testimony was a significant factor in passing these already popular measures; Sokal, 1990). Yerkes and Thorndike were members of a Eugenics Record Office committee on the inheritance of mental traits (Zenderland, 1998). In 1927, eugenics had become sufficiently well accepted that the Supreme Court ruled in favor of a Virginia law allowing for compulsory sterilization of individuals with intellectual disabilities. In writing the nearly unanimous decision, Oliver Wendell Holmes stated that "three generations of imbeciles are enough."

There can be little doubt that the decision to uphold compulsory sterilization and the restrictive immigration policies did irreparable harm and that intelligence testing and testers were closely aligned with these efforts. Nonetheless, it should be kept in mind that, at the time, these were popular positions. In addition to Oliver Wendell Holmes, the list of public figures who wrote in support of eugenics includes Winston Churchill, Margaret Sanger, and Helen Keller[15]—to name just a few. Views about eugenics changed dramatically after the Second World War when the Nazi plan to implement eugenics through mass extermination became public. Although intelligence testing continued, the link to heredity (and eugenics) received much less attention. It is difficult to know what the key figures in testing believed about eugenics in the second half of the 20th century; few advocated for it publicly.

Intelligence and Race

The controversy over intelligence testing came back with force in the late 1960s when Arthur Jensen began publishing on racial differences in intelligence. In 1969, *Harvard Educational Review* published a lengthy article by Jensen along with a series of invited responses from other psychologists. Jensen argued that the failure of compensatory education programs (such as Head Start) to produce lasting effects on IQ and achievement raised questions about the assumptions that supported development of these programs. In subsequent books and papers, Jensen further developed the idea of general intelligence (a view similar to Spearman's), argued

that a substantial proportion of the variance in IQ is the result of inheritance, and stated that the observed differences in IQ scores across racial groups cannot be explained by test bias.

Jensen's work attracted considerable attention both within the scientific community and with the general public, and he was accused of being a racist. Public speaking engagements were disrupted or canceled because of threats of disruption. At one point his university hired a bodyguard for him and his mail was routinely screened for bombs (Robinson & Wainer, 2006). Over the next 25 years, numerous books were published on both sides of the controversy.

The controversy over IQ, inheritance, and differences in racial and socioeconomic groups continued for over 2 decades. Jensen was not alone in advancing what was considered inflammatory research on IQ and race. He was also not alone in publishing outside academic journals in newspapers and magazines that were certain to produce a public response. Richard Herrnstein (1971) published in the *Atlantic Monthly* and Linda Gottfredson (1994) published in the *Wall Street Journal*. As with Jensen, publishing on this topic led to calls for removal from their respective universities. Gottfredson (Wainer & Robinson, 2009) lost funding for her position at Johns Hopkins shortly after she began this line of research. Gottfredson (Wainer & Robinson, 2009) and Jensen (1980) also clearly believed that they were excluded from some of the top-tier journals in their fields, not because of the quality of their research, but because of the controversial nature of their results. This may have motivated their decision to make the discussion not only public but also high profile by publishing outside the academic literature. Alternatively, it may be as Cronbach (1975) wrote:

> Scholars typically welcome the widest possible attention to their views because they cherish the ideas and because they prize visibility and influence. What wonder, then, that a scholar given his once-in-a-lifetime moment in the public eye will seek to make the moment memorable? (p. 12)

Although publication in this area has continued (e.g., Gottfredson, 2018), the last high-profile contribution to the public controversy came when Herrnstein and Charles Murray (1994) published *The Bell Curve: Intelligence and Class Structure in American Life*. This book became a *New York Times* bestseller and drew immediate attack within the mainstream press (Gould, 1996). Although this resurgence of the controversy no doubt sold many books and magazines, it is unlikely that it changed many minds. It does seem likely that the controversy created an unpleasant association between testing and racism and may have done more than a little to motivate the current antitesting movement.

In this section we saw how the development of the Army Alpha impacted the practice of assessment. In the next section we look more broadly at how other federal programs and legislation have impacted educational measurement.

THE INFLUENCE AND IMPACT OF U.S. GOVERNMENT INVOLVEMENT IN ASSESSMENT

As we discussed in the previous section, World War I provided a context for the introduction of large-scale standardized testing in the United States. The Army Alpha was, however, only one of the government-initiated programs that have impacted the theory and practice of testing. In this section we consider how both ongoing Department of Defense funding and federal legislation have continued to impact educational measurement.

It is not surprising that with the U.S. entrance into the Second World War the army would again turn to psychologists to help with the challenge of selection and placement created by the need to turn millions of enlisted men into a functioning army. By 1941, the idea of using tests for selection and placement in industry (and the military) was well established. There was an opportunity for psychologists to contribute to the war effort and at the same time advance the field. One of the most noteworthy individuals driving that effort was John Flanagan.

John Flanagan and the Aviation Psychology Program

Flanagan (1906–1996) had an impressive academic pedigree. As a high school science teacher, he attended summer psychology courses taught by E. L. Thorndike. He then left his teaching position in the late 1920s to study psychology at Harvard University, earning his degree under Truman Kelley. Flanagan then joined the Cooperative Test Service of the American Council on Education. During his time with the Cooperative Test Service, he produced one of the first significant expositions on test equating (Flanagan, 1939), which developed into the chapter on equating in the first edition of *Educational Measurement* (Flanagan, 1951). This chapter may well have been the first comprehensive presentation of the subject. Flanagan examined the importance of form development and comparable application of test specifications, as well as appropriate assignment of items to forms to support score comparability. He considered statistical approaches to equating as well as providing recommendations for the statistical machinery of equipercentile equating and arguing against the use of regression functions for equating because they provide an asymmetric link between test forms (Flanagan, 1951). Despite these significant contributions to the field, Flanagan is likely best known for his work on psychological testing during World War II, including development of the critical incident technique, which has been widely used in developing specifications for credentialing examinations and is still used in the early 21st century.

Like many of the other figures we have discussed, Flanagan's early work was tied to the eugenics movement. His initial experience in military research began in 1938 when he was hired by Frederick Osborn to conduct a study of the factors responsible for decisions regarding family size made by Army Air Corps officers and their wives (Flanagan,

1942). The foundation commissioning the study was concerned that Army Air Corps pilots—"considered the 'cream of the crop' genetically"—were having small families. The intention of the research was to identify "incentives to encourage them to have more children" (Capshew, 1999, p. 107).

This work led to Flanagan's commission in the U.S. Army Air Corps in 1941 to lead the Aviation Psychology Program. Under his leadership, more than 150 psychologists were similarly commissioned and contributed to the program, with more than 1,400 individuals contributing throughout the war ("APF Gold Medal Awards," 1993). The extensive work of the Aviation Psychology Program of the Army Air Forces is documented in 19 volumes describing studies in development and analysis of criterion measures, rating scales, checklists, and materials developed based on critical incident reports. The Aviation Psychology Program introduced criterion-referenced testing and test validation approaches based on performance. The validity research included comparisons of test results with data collected during training and subsequent performance in the field (Shields et al., 2001).

After working on the Osborn study, Flanagan was sponsored by the National Research Council to develop the Aviation Cadet Qualifying Examination, a test for classifying aircrew candidates. The examination was initially implemented shortly after the 1941 attack on Pearl Harbor (Staff, Psychological Branch, Office Air Surgeon, 1944). These tests were initially used for cadet selection and subsequently for placement into specific aircrew roles. The novel contribution introduced by Flanagan's team relied on the rapid development and validation of a complicated assessment battery. The assessment sorted participants into role-specific stanines for navigator, bombardier, and pilot roles, based on a combination of performance-based assessments (e.g., finger dexterity and rudder control), personality, and knowledge/proficiency tests. Validity evidence for the predictive power of these assessments was gathered through training measures (Did the individual successfully complete training?) and measures of success or failure in combat, including ratings, awards, errors, and survival (Flanagan, 1948).

A second research project, conducted during this period under the auspices of the Army Air Forces Aviation Psychology Program, represents a rarely possible textbook example of validity research for a selection procedure. This study was designed to evaluate the procedures and measures for selecting pilots and aircrew to maximize the match between their strengths and their assigned roles. The research reports of the Army Air Forces Aviation Psychology Program ("The Experimental Study of a Thousand Applicants Sent Into Pilot Training") described how approximately 1,050 men were assigned to pilot preflight school without consideration of their performance on the Army Air Forces qualifying examination battery. They were then enrolled in basic training and primary/advanced flying schools, with some eventually receiving ratings as qualified pilots. "Two-thirds of the men with stanine scores of 8 or 9 completed the program, while fewer than 3% of those with stanines 1 or 2 did so" (Jones, 2007, p. 604). Jones

also noted that the predictive value of the test was substantially reduced when applied to subsequent performance. The more than 20:1 ratio for completing training was reduced to less than 2:1 in a follow-up study predicting accidents associated with pilot error. And although follow-up during combat was plagued with problems arising from incomplete data, it was impossible to demonstrate that higher scores were associated with superior outcomes in combat.

Postwar Developments

The practical work that went on during the war continued to have a significant impact after the war ended. The formalization of the critical incident technique, an evidence-based test design approach used during the war to identify effective actions taken by officers, led to the development of the "critical requirements" of combat leadership for the U.S. Army Air Forces (Flanagan, 1954) and then to a set of procedures that have been broadly used in the construction of credentialing examinations. Robert Thorndike (1947, 1949) went on to devise methodologies for developing a personnel testing program based on his work in the Army Air Forces. These reports detail methods and considerations for performing job analysis, developing aptitude tests (pilot testing, validation, revision), criterion validation, developing composite scores from test batteries, and assessing reliability. Many of the individuals who participated in the Army Air Forces became leaders in the civilian research community—some, like Flanagan and Thorndike, were officers, whereas others, like William Angoff, participated as enlisted men (Jones, 2007).

Department of Defense Support for the Development of Modern Test Theory

Beginning during the Second World War and continuing through the 1990s, the Department of Defense supported a series of projects that advanced the development of what would become known as *item response theory*. The book *Measurement and Prediction* (Stouffer et al., 1950)—published after the war—includes a chapter on Paul Lazarsfeld's (1901–1976) wartime work on latent structure analysis. That work provided a basis for the conceptual and statistical development of IRT (e.g., Bock, 1997; Lord, 1952). In the 1950s, the U.S. Air Force School of Aviation Medicine supported Allan Birnbaum's development of the logistic form of the one-, two-, and three-parameter IRT models (Birnbaum, 1957, 1958a, 1958b). The Office of Naval Research supported the writing of Lord and Novick's (1968) volume that became a nearly sacred text for measurement experts and made Birnbaum's work widely available. The Office of Naval Research (along with other government agencies) also helped to move IRT from the theoretical to the practical by sponsoring a series of landmark conferences on latent trait theory and computerized adaptive testing. The conferences took place in 1975 (Clark, 1976), 1977 (Weiss, 1978), and 1979 (Weiss, 1983). These conferences represented the state of the art in what became known as IRT and the presenters represented a who's who of researchers working in that area. The authors listed in these proceedings include Fumiko

Samejima, Frederic Lord, Bert Green, Darrell Bock, Dan Eignor, Ron Hambleton, Mark Reckase, H. Swaminathan, Linda Cook, David Thissen, and Howard Wainer. The Department of Defense also sponsored a more than 30-year-long research project that resulted in computerization of the Armed Services Vocational Aptitude Battery (ASVAB) in 1996. The foreword to a report on this effort described this accomplishment (Sands et al., 1999).

> In October 1996, the Department of Defense (DoD) implemented a computerized adaptive testing (CAT) version of its enlistment test battery (the Armed Services Vocational Aptitude Battery or ASVAB) in 65 Military Entrance Processing Stations (MEPSs) across the country. DoD became the first organization to use CAT-derived scores for personnel selection when the system was placed in five MEPSs for operational testing in 1992; now DoD has become the first employer to adopt CAT for its employment system. . . . The Department is the largest single employer of American youth, testing over 350,000 applicants for entrance into the Military Services between October 1, 1994 and September 30, 1995. (p. ix)

The ASVAB project began in the 1960s. Early aspects of this work focused on statistical models for item selection, but subsequent work covered every aspect of implementation, including pool construction, exposure control, software and hardware selection, and validity studies. This effort produced important advances in computerized adaptive testing, but more generally it advanced the field with respect to both technical and practical knowledge of IRT. As important as these contributions were in the evolution of educational measurement, they may well be judged as having a minor impact when compared to the other way the federal government influenced the measurement field—through legislation to support education and assessment.

The Impact of Legislation on Educational Measurement

In response to the Soviet launch of *Sputnik* in 1957 and the perception that America was losing the space race, federal policy began to explicitly support testing of schoolchildren in science, mathematics, reading, and foreign languages through the 1958 National Defense Education Act (NDEA). This served the purposes of establishing baseline knowledge in these areas as well as providing support for rapid curriculum development and implementation in public schools (Bunch, 2021). The testing programs supported by the NDEA were designed to evaluate college readiness, specifically preparedness for postsecondary study in math, science, and foreign languages. The act also provided federal funding for these tests in the event that a state could not legally pay for them (Section 504(b)).

Approaching federal involvement in education as a matter of national defense distinguished the NDEA from previous proposed education-spending legislation, versions of which had been passed by the Senate but languished in the House. Framing support for standard educational goals (and assessments) across states as a matter of national defense was ultimately a winning strategy, although there was controversy about the

appropriate limits and uses of standardized testing in schools. Advocates for this legislation saw the potential to provide greater access to college through federally supported scholarships, loans, and selection instruments, while challengers were skeptical about the utility of standardized tests in making decisions about individual fitness for secondary education or a career path (Urban, 2010).

The programs underwritten by the NDEA continued through the 1960s and eventually were subsumed by the Elementary and Secondary Education Act (ESEA) in 1965. The ESEA was a part of Lyndon Johnson's Great Society initiative and laid the groundwork for state and school district accountability for educating disadvantaged children. The ESEA was reauthorized and/or amended regularly through 2015 (taking on names including No Child Left Behind and Every Student Succeeds Act). The ESEA also brought the federal government into the education of bilingual students and students with disabilities, provided federal money for testing and supplemental educational services for historically disadvantaged students, and established Head Start as a permanent program (Bunch, 2021; Urban, 2010).

A significant part of the motivation to increase the assessment component of this legislation was to ensure that states and localities would effectively implement the requirements for the education of disadvantaged students and students with disabilities. The assessment component was again strengthened with the passage of the 1974 amendments (Education Amendments, 1974) requiring the commissioner of education to develop program evaluation models that "specify objective criteria" and provide "comparable [data] on a statewide and nationwide basis" (Section 151; Bunch, 2021). During the 1970s and 1980s, states were supported by 10 regional technical assistance centers to develop and implement assessments in support of ESEA requirements. These centers employed hundreds of testing specialists and interacted with education officials in all 50 states to create a standardized approach to solving educational measurement problems. Between 1979 and 1981, these centers provided over 5,000 workshops and on-site consultations for more than 80,000 clients (Bunch, 2021; Stonehill & Anderson, 1982).

In addition to promoting assessment through these regional centers, the 1978 reauthorization directed the commissioner of education to assist state education agencies in developing capacity to conduct large-scale achievement testing programs to measure reading, writing, and mathematical proficiency, along with other subjects. After the election of Ronald Reagan in 1980, the ESEA was reauthorized as the Educational Consolidation and Improvement Act. This led to more local control, but somewhat surprisingly it again increased the requirements for assessment. As a result of this act, schools, districts, and states were required to administer criterion-referenced tests to demonstrate that targeted outcomes were being met.

Despite the years of legislation and associated funding, the 1983 report on the state of American education, *A Nation at Risk* (National Commission on Excellence in Education, 1983), described America's schools as failing to educate the nation's children. This led to the 1994 Improving America's Schools Act. The act led to additional assessment requirements: Assessments aligned with the content standards were to be administered

at a point between Grades 3 and 5, between Grades 6 and 9, and again between Grades 10 and 12 (Section 1111(b)(3)(D)). Standards, however, were still established at the state level and were highly variable across states.

In 2002, this trend of increasing emphasis on assessment continued with the passage of No Child Left Behind (NCLB), in place from 2002 to 2015. The act expanded testing to include Grades 3 through 8 and required testing in one grade of high school. As with every reauthorization of the ESEA since 1966, NCLB emphasized the importance of education and testing of minority populations, English language learners, and students with disabilities. NCLB created an emphasis on state standards and assessment that led to meaningful developments in federal education policy. Building on the goal of all children achieving proficiency in reading and mathematics (by the 2013–2014 academic year), NCLB introduced test-based accountability provisions such as annual improvement targets and penalties if targets were not met. NCLB accountability requirements increased policy attention on test accommodations and accessibility, as well as expectations for progress reporting for students and student subgroups.

There can be little doubt that federal legislation has done much to drive assessment of specific populations, influence test accommodations and tested content, and impact how testing is performed in education for students in kindergarten through Grade 12. The legislation described has also shaped the significant role that measurement experts have had in education and accountability efforts since the 1960s. As significant as this has been, it may be that even more psychometric innovation was driven by a different federal initiative, NAEP. We would be remiss not to mention NAEP in this section on federal involvement in assessment, but we will present the details of this program in the subsequent section on large-scale testing.

RELIABILITY THEORY AFTER SPEARMAN

In the prior two sections we focused on the historical and social developments that shaped educational measurement. In this section we return to the more theoretical aspects of the history of educational measurement and follow the development of reliability theory after Spearman's early papers. This development has two parts that are sometimes viewed as chronologically distinct but are, in fact, interwoven. The first has to do with the continued development of classical test theory; the second part resulted in generalizability theory.

The Development of Classical Test Theory

Although Spearman's work provided the foundation for classical test theory, his focus was on creating tools for correlational psychology. He was interested in efficient approaches to estimating correlations between tests and in tools for evaluating those relations. There is little evidence that his interest extended to advancing methodology

for estimating standard errors or otherwise interpreting individual test scores. This was consistent with his interest in using test scores to study intelligence rather than to evaluate individuals. Spearman was a correlational psychologist, not a psychometrician. In this respect, Spearman had a great deal in common with Galton and Pearson. All three were interested in developing correlational tools to advance their respective substantive fields: heredity, evolutionary biology, and human intelligence. The next major contributor to the development of test theory, Truman Kelley, approached testing from a different perspective: that of a methodologist. In fact, Cronbach referred to Kelley as an "obsessive algebraist and statistician . . . who had no motive save methodologic" (Lee Cronbach, personal communication, December 22, 2000).

Truman Kelley

Truman Kelley (1884–1961) earned his doctorate under E. L. Thorndike at Columbia and went on to teach at Stanford and Harvard. In 1922–1923, Kelley spent a sabbatical year with Pearson at the Galton Biometric Laboratory. Around this same time, Kelley began publishing a series of papers that would extend classical test theory beyond the formulas presented by Spearman. What may be Kelley's single greatest contribution to classical test theory was among his first, the regression estimate of a test taker's true score:

$$\hat{T}_x = \rho X_i + (1 - \rho)\bar{X}. \tag{5}$$

In this equation taken from Kelley (1923), the left-hand term represents the estimated true score, ρ represents the reliability of the test score, X_i is the observed score for test taker i, and $\bar{X}$ is the population mean. This equation not only represents an important use of the implicit assumptions of classical test theory, but also links test theory to a Bayesian framework by combining the observed data with a prior assumption (Levy & Mislevy, 2021).

The next year, he published a paper showing that if we have two forms of a test such that $x_1 = a + e_1$ and $x_2 = a + e_2$, where x is the observed score, a is the true score, and e is a "chance factor,"[16]

$$r_{12} = \frac{\sigma_a^2}{\sigma^2}. \tag{6}$$

In Kelley (1925) he showed the relationship between the reliability of a measure and the correlation between the measure and the related true score:

$$r_{1\infty} = \sqrt{r_{1I}} \tag{7}$$

or, equivalently,

$$r_{1I} = r_{1\infty}^2. \tag{8}$$

These equations will be familiar to anyone who has studied classical test theory; they represent only a sample of Kelley's contribution. In addition to numerous additional papers of this sort, Kelley published what may well be the first book devoted to the mathematical theory of test scores, *Interpretation of Educational Measurements* (1927), and substantially influenced formal presentations of test theory that followed. The extent to which Thurstone's *The Reliability and Validity of Tests* (1931) and Gulliksen's *Theory of Mental Tests* (1950) depend on Kelley's work makes it clear just how essential Kelley's contributions were to the development of educational measurement.

G. F. Kuder and M. W. Richardson

G. F. Kuder (1903–2000) and M. W. Richardson's (1896–1965) 1937 paper followed in the spirit of Kelley in showing how simple assumptions and algebraic manipulation can be used to derive practically useful results. Their paper began with a theoretical discussion of reliability and then progressed through a series of manipulations to produce what have become known as the KR20 and KR21 (the 20th and 21st equations presented in the paper). The first of these represents a generalized equation for reliability and the latter a form that the authors stated could be calculated in 2 minutes using statistics that are readily available to the test developer. Their presentation began with a rejection of the test–retest paradigm for reliability in favor of the split-half approach (highlighting the importance of the Spearman–Brown formula). They noted that split-half methods depend on the specific split used and then presented a general equation that represents the correlation between two equivalent forms of a test. If $a, b, \ldots n$ and $A, B, \ldots N$ are corresponding items in two hypothetical forms of a test, the tests are equivalent if the items in each pair (a and A, b and B, etc.) are interchangeable. Based on this assumption, they present the KR3:

$$r_{tt} = \frac{\sigma_t^2 - \sum_1^n pq + \sum_1^n r_{ii} pq}{\sigma_t^2}. \tag{9}$$

In this equation, r_{tt} is the reliability of interest, σ_t^2 is the variance of the test, p is the difficulty for a specific item, and q equals $1 - p$. They recognized that this formulation was not of practical use because r_{ii} cannot be calculated. After 16 intervening equations, they presented the KR20, derived from KR3 with the assumption that all intercorrelations are equal:

$$r_{tt} = \frac{n}{n-1} \cdot \frac{\sigma_t^2 - n\bar{p}\bar{q}}{\sigma_t^2}. \tag{10}$$

Finally, by assuming that all item difficulties are equal, they produced the KR21,

$$r_{tt} = \frac{n}{n-1} \cdot \frac{\sigma_t^2 - n\bar{p}\,\bar{q}}{\sigma_t^2}. \tag{11}$$

Kuder and Richardson's (1937) paper was important both practically and theoretically. For practitioners, it provided an approach to estimating the reliability of a test that was computationally simple (at the cost of making strong assumptions). In the days before computers and electronic calculators, this was of considerable value. Their paper also represented an important theoretical step forward in the understanding of reliability.

Lee Cronbach

A final critical step in the development of reliability within the framework of classical test theory was provided by Cronbach's introduction of coefficient alpha. Cronbach (1916–2001) made major contributions to assessment. In addition to his work on advancing the conceptualization of reliability theory, he published groundbreaking work on validity theory. He coauthored the seminal paper on construct validity with Paul Meehl (Cronbach & Meehl, 1955) and wrote the chapter on validity for the 1971 edition of *Educational Measurement.* As an interesting aside, Cronbach was also a member of Terman's gifted group. The gifted group was followed long after Terman's death, and in addition to being a subject in the study, Cronbach assisted with production of the final volume of the study (Holahan et al., 1995).

Cronbach's 1951 paper on alpha may be the most cited paper in the history of measurement. According to Google Scholar, it has been cited tens of thousands of times. In a sense, the coefficient alpha paper extended the presentation from Kuder and Richardson (1937). KR20 is a special case of coefficient alpha in that it applies to items scored 0/1. Cronbach showed that coefficient alpha (and, by extension, KR20) represents the mean of all possible split-half coefficients.[17] It is also "the value expected when two *random* samples of items from a pool like those in the given test are correlated" (Cronbach, 1951; p. 331).

In a posthumously published paper, Cronbach reflected on coefficient alpha (Cronbach & Shavelson, 2004). He argued that although reliability coefficients were appropriate for the kind of correlational psychology that was prevalent from Spearman's time through the first half of the 20th century, more recent applications of measurement are better served with an estimate of the standard error of measurement.[18] In that same paper, Cronbach advocated for procedures based on analysis using variance components rather than correlations. He was arguing that generalizability theory should supplant classical test theory as the preferred framework for conceptualizing reliability.

Generalizability Theory

Cronbach's advocacy for the framework represented by generalizability theory comes as no surprise; he spent over a decade working through the problem of reliability before he and his colleagues published *The Dependability of Behavioral Measurements* (Cronbach et al., 1963, 1972). The work created what the authors referred to as a liberalization of classical test theory.

In classical test theory, observed scores are conceptualized as comprising a true score and an undifferentiated error term. The true score is the expected score that the test taker would receive if they could complete the test an unlimited number of

times. In generalizability theory, the sampling process is conceptualized as occurring at the task (or item) level, with tasks being randomly sampled from a universe of admissible tasks. This allows for multiple facets comprising the total score variance to contribute either to the test taker's universe score (similar to the true score) or to error. The framework allows for conceptualizing the reliability of scores in which (a) separate components of the error variance are associated with the sampling of the items and the raters who evaluate the responses to those items, (b) the total score is a composite based on stratified sampling of items from different content areas or item formats, (c) the measurement is focused on scores for individual test takers or aggregates such as classrooms, and (d) relative versus absolute score interpretations are possible.[19]

Clearly, any history of generalizability theory will need to include the work of Cronbach and his collaborators beginning in the late 1950s and lasting very nearly until Cronbach's death, but many of the ideas reflected in generalizability theory first appeared in the late 1800s, and as with classical test theory, they were developed by numerous individuals over time. The earliest of those ideas came from the work of Francis Edgeworth (1845–1926).

Edgeworth attended Trinity College, Dublin, and Oxford University, where he studied ancient and modern languages. He was self-taught in the areas in which he made his reputation: statistics and economics. His work in economics earned him a chair at Oxford. Edgeworth was interested in bringing mathematical analysis into the social sciences. This interest led him to become one of Galton's earliest disciples (Kendall, 1968; Stigler, 1986). This interest also led him to write three papers on the statistics of examinations. Edgeworth's interest expressed in these papers was to examine "the degree of accuracy or inaccuracy which is to be ascribed to the modern method of estimating proficiency by means of numerical marks" (Edgeworth, 1888, p. 600); in other words, he was interested in the reliability of test scores.

When Edgeworth referred to the "modern method," he was speaking of a system in which essays or short-answer papers were scored by content experts. The purpose of the paper was to describe an approach for estimating the extent to which chance influenced examination scores. He began by describing something closely related to what we now think of as the true score in classical test theory (or the universe score in generalizability theory). Using Latin translation as an example, he wrote,

> This central figure which is . . . assigned by the greatest number of equally competent judges, is to be regarded as the true value of the Latin prose; just as the true weight of a body is determined by taking the mean of several discrepant measurements. There is indeed this difference between the two species of measurement, that in the case of material weight we can verify the operation. We can appeal from kitchen scales to an atomic balance, and show that the mean of a great number of rough operations tends to coincide with the value determined by the more accurate method. But in intellectual ponderation we have no atomic

> balance. We must be content without verification to take the mean of a number of rough measurements as the true value. (Edgeworth, 1888, p. 601)

He went on to define error in terms of variability about this true value and concluded that for aggregate scores error "will fluctuate according to the normal law" (p. 604). Having described the observed score as being composed of a true score and a normally distributed error, Edgeworth then presented a number of the essentials of generalizability theory:

1. He described the total error score as being made up of multiple facets, including what in generalizability theory would be designated as task effects, rater effects, and rater-by-task interactions.
2. He made the assumption that tests were constructed through random sampling from a defined domain—the central assumption of generalizability theory (Kane, 2002)—and commented on conditions that would violate the model.
3. He described the total error as the square root of the sum of the squared effects and noted that the impact of any effect will be reduced by the square root of the number of observations contributing to that effect.

Edgeworth used these conceptualizations to describe data collection designs that would minimize the error without increasing the total amount of rater time required.

Having considered sources of error associated with raters, Edgeworth (1888) then examined the contribution of task sampling to the error of measurement:

> We have so far been endeavouring to estimate the error which is incurred in appreciating the actual work done by the candidate. We have now to evaluate the error which is committed in taking his answers as representative of his proficiency. (p. 614)

The suggestion that Edgeworth presented the basics of generalizability theory in this paper may seem like an exaggeration because generalizability theory uses the framework of analysis of variance and R. A. Fisher did not publish his ideas until decades after Edgeworth's paper (e.g., R. A. Fisher, 1925). It is, nonetheless, obvious that this is the framework Edgeworth intended; he had already published a paper on "analysis of variance" in 1885. Stigler (1986) commented on the 1885 paper that "the solution was insightful and foreshadowed much of twentieth-century work on the analysis of variance" (p. 314).[20]

Edgeworth's insights into the generalizability of test scores were impressive. Whether they influenced subsequent work in this area is difficult to say. It may be that, like his contribution on analysis of variance, these papers were largely ignored—perhaps he was too far ahead of his time. We have found no references to his work in the early literature on test theory, but this is questionable evidence because extensive citation of previous work was rare at that time. For most of the half century that followed Edgeworth's 1888 paper, the emphasis was on the development of classical test theory, which has

already been discussed. The next important contributions to reliability theory that went beyond the classical test theory paradigm were provided in 1936 by Harold Gulliksen (1903–1996) and Cyril Burt (1883–1972).

Gulliksen's (1936) contribution appeared in the first volume of *Psychometrika.* He looked at the typical approach to reliability for objectively scored items; because variability in this context is substantially associated with item sampling, he referred to it as content reliability. He then raised the question of how content reliability can be estimated for essay examinations in which the total error includes both content error and error introduced by "readers." Assuming that two separate forms of the test are available and that one of those forms is scored by two separate readers, he developed an approach to quantifying the content reliability associated with essay tasks.

Burt's (1936) contribution is framed in the context of factor analysis. In his publication, he conceptualized total score as being constructed of multiple factors, including a general factor and sources specific to raters. Using factor analytic and correlational procedures, he described how rater error can be quantified.

Cronbach et al. (1972) described the contributions by Gulliksen (1936) and Burt (1936) as being among the first to conceptualize measurement error as comprising multiple facets. But as Cronbach (1991) would later write, univariate generalizability theory "interweaves ideas from at least two dozen authors" (p. 394). These contributions include R. A. Fisher's work on analysis of variance (R. A. Fisher, 1925), as well as that of Cornfield and Tukey (1956). Burt (1955) also made a second contribution that explicitly linked estimation of reliability to analysis of variance.

These publications and numerous others set the stage for the work Cronbach and his collaborators did in the late 1950s and 1960s. In 1957, Cronbach and Gleser began work on a handbook of measurement theory. Believing that reliability had been thoroughly studied by others, they decided to begin with that section of the book. Cronbach (1991) wrote later, "We learned humility the hard way—the enterprise never got past that topic" (p. 392).

The essential features of univariate generalizability theory were largely completed by the early 1960s and published in three journal articles, each with a different first author (Cronbach et al., 1963; Gleser et al., 1965; and Rajaratnam et al., 1965). These papers incorporated analysis of variance procedures to evaluate multiple facets representing different sources of error, absolute versus relative error, and an understanding of fixed versus random facets in the design. Additional years of effort were then necessary to expand the theory to a multivariate framework. The result was *The Dependability of Behavioral Measurements* (Cronbach et al., 1972).

Their monograph represents a remarkable step forward in both the conceptualization and the estimation of measurement error. Unfortunately, although the volume captured much of the authors' best thinking on the topic, as Cronbach (1991) wrote, "the book proved indigestible" (p. 395). Cronbach et al. (1972) warned that "the reader is certain to gain far more from his third reading of most sections than from his first or second" (p. 4). Although Cronbach continued to work in this area until shortly before his death,

he published relatively little on the topic after the 1972 volume. The extent to which generalizability theory has become widespread seems to a substantial degree to be the result of subsequent authors who have developed the theory and written texts (and software) that make it more accessible. Although numerous researchers have written on the topic, far and away the greatest contribution has come from Robert Brennan. For many years, *Elements of Generalizability Theory* (Brennan, 1983) was the primary introduction that researchers had to generalizability theory. This monograph provided the basics needed for univariate generalizability analysis and prepared readers to make use of *The Dependability of Behavioral Measurements* (Cronbach et al., 1972). Brennan's more complete text and the associated software (Brennan, 2001a, 2001b) provided a detailed resource for multivariate generalizability analysis.

In concluding, we wish to make clear that we do not intend to imply that the technology for evaluating reliability and estimating measurement error is now complete. Instead, our goal has been to explain how this technology has developed during the past century. This summary is intended to describe how these concepts have evolved. Hopefully, this will lead to a better understanding of the current state of the art and will at the same time point to areas for development of both theory and practice.

SCALING

Next, we consider the development of applications of scaling for educational measurement. Unlike the other sections of this chapter, for which the content is likely to be self-evident given the section heading, the term scaling may warrant a definition, or at least a description. This is not because it is an obscure term, but because a focus on scaling has been in and out of vogue over the decades, and ideas about the requirements for a defensible scale have varied as well. In 1954, Lord wrote that "'scaling' appears to be virtually indistinguishable in meaning from 'measurement,' which may be defined as the assignment of numerals to objects (including people), according to some rule, in order to represent their properties" (p. 375). Lord went on to state, "If a method of scaling rests upon a verifiable hypothesis as to the existence of some specified self-consistency in the data, then the data may or may not be found to scale" (p. 375). Lord's first statement sets a low bar for scaling and, by extension, for measurement. The second implies that if the user establishes a higher bar by way of the intended interpretation of the scores, it may be necessary to demonstrate that the data—and, by extension, the phenomenon represented by those data—support that interpretation. Lord's statement suggests that scaling has occurred even when it is strictly categorical; scales need not be continuous, nor do they need to represent ratio or even interval measurements.[21]

The first two editions of *Educational Measurement* included chapters on the nature of measurement. Consistent with much theory and practice at the time, this chapter did not appear in the third and fourth editions. More recently, the limited attention that has been given to the theoretical requirements of measurement (and scaling) has become a source of some controversy. Michell (1999) has written extensively and crit-

ically on this topic. McGrane and Maul (2020) described the failure to demonstrate that the latent traits reflected in IRT models have characteristics that support the use of the associated scales as a "scientifically dire situation" (p. 8). Recently, several authors have attempted to resolve this controversy by placing scaling procedures in perspective. Briggs (2021) provided a historical perspective incorporating the contributions of Fechner, Galton, Thurstone, and Stanley Stevens. Mari et al. (2021), in their volume *Measurement Across the Sciences*, tried to define the requirements of measurement in psychology from a broader scientific perspective and specifically by bringing in the perspectives of metrology. David Torres Irribarra's (2021) volume *A Pragmatic Perspective of Measurement* provides yet another view of the controversy bringing both philosophical pragmatism and common-sense pragmatism to the discussion.

For this brief summary of the history of scaling, we focus on developments in practice. We recognize that separating developments in the taxonomy of scaling procedures (see Stevens, 1946, 1951; Torgerson, 1958) and the conceptualization of the nature of measurement more broadly from developments in the practice of scaling is somewhat artificial, but these theoretical issues are presented in detail in Briggs et al. (this volume). To outline the development of scaling, we return to Galton and Fechner, but this time we focus on their efforts to build measures. We then examine how these early efforts at scale development impacted the development of the IRT models that have become ubiquitous. (See Moses, this volume, for a presentation of current approaches to scaling.)

Ernst Weber and Gustav Fechner

In 1860, Gustav Fechner provided an initial formalization of psychological scaling with the publication of *Elements of Psychophysics* (1860/1966). As we described previously, Fechner's magnum opus was the culmination of a decade of work conducting experiments on individuals' responses to physical stimuli and extending the work of Ernst Weber. Weber had spent his career examining psychological responses to physical stimuli and had found that when asked to detect differences in physical stimuli, subjects were sensitive to relative changes in proportion, not absolute changes in magnitude. Fechner generalized this work by recognizing that the sensory intensity of a stimulus is the product of the log of the level of physical stimulus multiplied by some constant K(Sensation = K log Stimulus).[22] This finding, later described as Fechner's law, indicates that there is a logarithmic relationship between a physical stimulus and the associated perception. That is, if you need to double the stimulus to increase the perceived difference from 1 to 2, you must double it again to increase the perceived difference from 2 to 3 (Briggs, 2021).

Although Weber had primarily concerned himself with comparing the magnitude of these proportions across different sensory stimuli, Fechner recognized that these psychological attributes could be expressed on their own scale separate from the scale of the stimulus itself. Fechner noticed in his experiments that subjects sometimes disagreed as to which stimulus was the strongest. He hypothesized that these disagreements were attributable to random error in the assessment of the stimuli. He further hypothesized that these errors in measurement were normally distributed with a mean of zero.

Therefore, subjects disagreed as to which stimuli were more intense because the psychological comparison on the stimuli follow two random variables, each following a normal distribution (Fechner, 1860/1966).

In his experiments, Fechner compared two physical stimuli across many subjects and calculated the proportion of time that the treatment stimulus was judged to be greater than the control stimulus. Then, assuming that the standard deviations of these responses were the same for all pairs and stimuli, Fechner set these equal to 1 and used the inverse of the cumulative density function to compute an estimate of the mean difference in psychological response for those stimuli. After repeating this across multiple pairings of stimuli, Fechner was able to establish a scale for the measurement of sensory intensity.

Fechner's work was important in shaping thinking about the nature of scaling and measurement in psychology—and, by extension, educational measurement—because it established an early precedent that such measurement should follow the framework used in the physical sciences. This goes well beyond Lord's "assignment of numerals to objects . . . according to some rule" (Lord, 1954, p. 375).

Francis Galton

Although Fechner's work is notable as the first attempt to scale psychological stimuli, his work focused on the scaling of stimuli, not subjects. The first significant advance in scaling subjects was achieved through the work of Francis Galton. Although, as we have discussed, Galton is better recognized for his research on inheritance and his discovery of correlation, his work in scaling unobservable human characteristics provided an important foundation for the scaling of mental tests that is still in use in the early 21st century.

Inspired by the work of Adolphe Quetelet, Galton had become fascinated with the normal distribution and its ubiquity in nature (Briggs, 2021; Gillham, 2001). Galton recognized that the distribution of some physical traits, for example, height, was the product of many small, random events in the inheritances of each individual. He believed that unobservable mental traits were equally a product of this random path of inheritances and reasoned that they too should follow a normal distribution. This revelation regarding the shape of the distribution allowed Galton to translate the measure of any individual into its percentile—a word Galton coined—allowing individuals to be placed relative to one another along a statistical scale (Galton, 1875).

Galton referred to this approach to scaling as "relative" and recognized that although it allowed human qualities to be quantified, it lacked an objective reference point. In an effort to ameliorate this limitation, Galton developed scale anchor descriptors to indicate the performance that would be expected at particular percentiles. Furthermore, Galton recognized that relative assessments could be every bit as useful as absolute measurement in the myriad situations where individuals are assessed relative to a peer group, saying that "a blurred vision would be above all price to an individual man in a nation of blind men, though it would hardly enable him to earn his bread elsewhere"

(Galton (1889), p. 36). This relative approach to scaling individuals was valuable for the study of inheritance; by the turn of the 20th century, it was about to make its way into educational measurement. Again, we see physical measurement as the model for psychological measurement.

E. L. Thorndike

The next important contribution to scaling—and the first that was consciously intended to advance the field of educational measurement—was made by E. L. Thorndike (1874–1949). In the early 20th century, secondary education was becoming widespread in the United States. Schools all over the nation needed teachers, and increasingly, large urban universities were adding departments of education to address this need. Columbia University was at the center of this trend, and in 1899 Thorndike became an assistant professor at Columbia Teachers College (Cherry, 2023). Thorndike believed that science was the only sure foundation for social progress and was unhappy with the level of scientific rigor applied to education at that time (Clifford, n.d.). He admired the work of Galton and Pearson and wished to apply the scientific method to the study of education. To this end, he introduced the first course in educational measurement in 1902 and published a textbook, *Educational Psychology*, in 1913.

During his 40 years at Columbia, Thorndike worked on the design and development of many exams for a wide array of purposes (including his evaluation of the Army Alpha described previously). Trained in psychology, Thorndike had limited formal training in mathematics and statistics. Nonetheless, he built on the work of Galton when developing a scaling technique that placed students across grades on a common scale. To accomplish this, Thorndike began by calculating the proportion of test takers answering each question correctly for each grade. Items were then expressed as a deviation from the grade mean in terms of standard deviation units. The mean difference between common items administered in adjacent grade levels was then calculated and used to adjust grade means to place all grades on a common scale (Thorndike, 1919). Although this represented a significant advancement in educational scaling in 1905, this method assumed that the distribution of abilities across grades differed by position but not by dispersion (Engelhard, 1984; Thurstone, 1927). Twenty-five years later, this limitation would be addressed by Thurstone.

L. L. Thurstone

The year that Gustav Fechner passed away in Leipzig, L. L. Thurstone (1887–1955) was born in Chicago, Illinois. While at Cornell studying engineering, Thurstone's interest in education and psychology grew, along with his frustration with the poor quality of teaching in the college of engineering. In 1914, Thurstone abandoned engineering and matriculated at the University of Chicago to pursue his doctorate in psychology (D. A. Wood, 1962). During his graduate training, he accepted an assistantship at the newly created Division of Applied Psychology at the Carnegie Institute of Technology. During the next 9 years at Carnegie, Thurstone was at the forefront of the development

of psychological tests on a wide variety of cognitive functions. In 1924, he returned to the psychology department at the University of Chicago, where he taught a course on test theory.

At this point in his career, Thurstone began to scrutinize the scaling approaches that were common in mental assessments at the time, like those used by Thorndike and others. Thurstone recognized that earlier approaches to educational scaling assumed that dispersion was constant across grade. Thurstone believed this assumption was dubious and in 1924 he articulated a method of scaling that accounts for differences in both group means and dispersion (Thurstone, 1927). Thurstone's absolute scaling approach begins similarly to Thorndike's approach. For each test question, the difficulty is calculated for students in that grade and those difficulty values are rescaled to be expressed as deviations from the group mean in standard deviation units. Unlike Thorndike, however, in the absolute scaling approach, Thurstone calculated that standard deviation of the difficulty for common items in adjacent grades. The ratio of these standard deviations could then be used to rescale one grade onto the scale of the other grade while accounting for the differences in dispersion. This process could be repeated across adjacent grades to create a single scale that spanned several grade levels.

In 1925, Thurstone extended this scaling work to consider placing test takers along these educational scales. To unify the scales for test takers and items, Thurstone placed each test question on the test-taker ability distribution so that the percentage of test takers answering the question correctly was equal to the percentage of test takers to the right of that point in the distribution. This allowed item difficulties to be expressed as the place along the test takers' ability distribution at which 50% of test takers would answer the item correctly. He then represented these item difficulties in terms of standard deviation units, such that items at the mean ability were scaled to zero while more difficult and easier items were expressed as positive and negative normal deviates, respectively. Thurstone went on to graph the empirical probability of a correct response across grade for items at different difficulties and showed that the difficulty distributions are S-shaped, with slopes that vary by item. He stipulated,

> A refined statistical method for ascertaining the age at par [item difficulty] for each test question would be to fit an equation to the curve for each test question and then to ascertain the point at which the curve intersects the 50% level. (Thurstone, 1925, p. 446)

Thurstone did not develop that refined statistical method. As we will see in the next section, that development would result in the creation of IRT.

The specific examples of scaling that we provided in this section represent major steps in the development of scaling procedures in educational measurement. Each of these procedures goes well beyond Lord's simple assignment of numerals based on a rule; each assumes "some specified self-consistency in the data" (Lord, 1954, p. 375) that contributes

to the interpretation of the measurements. As the works cited earlier in this section make clear (Briggs, 2022; Irribarra, 2021; Mari et al., 2021; McGrane & Maul, 2020), the defensibility of these types of interpretations remains controversial. The historical perspective provided in this section does not resolve that controversy, but it does provide a context for understanding it. The practical application of these efforts to construct self-consistent scales resulted in IRT—which we consider in the next section—but as we discuss in that section, it did not eliminate the controversy over the requirements for scaling.

ITEM RESPONSE THEORY

The history of IRT is complex. Part of that complexity comes from the fact that IRT as we know it in the early 21st century developed independently—at about the same time—in both the United States and Denmark. In Denmark, that work was the effort of Georg Rasch. In the United States, numerous researchers made individual contributions that supported the ultimate development of IRT.

The Foundations of IRT

Development of the One-, Two-, and Three-Parameter Logistic Models

The lineage of IRT connects it both to the work on classical test theory described earlier in this chapter and to the work on scaling described in the previous section. In 1934, the United States was in the grips of the Great Depression, and the Roosevelt administration was committed to putting Americans back to work through programs like the Works Progress Administration. During this time, Thurstone was teaching psychology at the University of Chicago when the government offered him a new research assistant, Ledyard Tucker (1910–2004; Carlson & von Davier, 2017). Like Thurstone, Tucker had an undergraduate degree in electrical engineering; Thurstone quickly recognized Tucker's talents and encouraged him to continue his education in psychology. After more than a decade collaborating with Thurstone, Tucker earned his doctorate in 1946 and took a position at the College Entrance Examination Board (College Board; Dorans, 2004).

During his brief time at the College Board, Tucker defined the mathematical form for what we now refer to as an item characteristic curve (ICC). Unlike Thurstone, who had calculated nonparametric ICCs using observed data, Tucker used the integral of the normal curve—the normal ogive—to relate test-taker ability to performance (Tucker, 1946). This work made it possible to calculate the probability that a test taker at any ability level would answer a particular item correctly.

Although Tucker described the mathematical form for the ICC, he did not devote much attention to the underlying latent trait, instead relying on the language of true score theory to describe test-taker ability. This next critical step was based on the work of Paul Lazarsfeld (1901–1976), who in his 1950 publication laid the conceptual foundation for the *latent trait*, in what psychometricians soon would call latent trait theory.

Lazarsfeld studied mathematics in Vienna. With the rise of Hitler, he left Austria for the United States and eventually moved to Columbia University, where he became the associate director of the university's Bureau of Applied Social Research ("Dr. Paul Lazarsfeld Dies," 1976). Although Lazarsfeld is known in sociology for his work on voting and the media, his contributions to test theory stem from his development of latent structure analysis. Latent structure analysis posits that there is an unobserved, *latent* cause for a subject's observed responses. By assuming that responses to individual items are independent, Lazarsfeld was able to study the joint occurrences of item response patterns to make inferences about the trait that underlies this behavior (Lazarsfeld, 1950). Although, the majority of Lazarsfeld's work concerned attitude measures, not educational assessments, the reconceptualization of observed data as being the product of a latent trait as opposed to the trait itself was a critical development.

Bert Green (1928–2019) recognized the applicability of Lazarsfeld's latent structure analysis to educational measurement. In the early 1950s, Green set to work analyzing test-taker response data using latent structure and latent class models. He extended Lazarsfeld's work by developing a general procedure for latent structure and latent class models (Green, 1951a, 1951b) and demonstrating the interrelationship between factor analysis and latent structure analysis. Through this work, Green directly connected latent structure analysis models to what would become IRT. The final steps in the introduction of a statistically rigorous theory of item response patterns was then provided by Lord (1952) and Birnbaum (1957, 1958a, 1958b).

In 1944, Frederic Lord (1912–2000) joined the graduate record office of the Carnegie Foundation, the forerunner to Educational Testing Service (now known as ETS). In 1949, he became the director of statistical analysis at the newly formed Educational Testing Service (Carlson & von Davier, 2017). He completed his dissertation in March 1951, and in 1952 the work was published as a psychometric monograph. The monograph presented the one- and two-parameter IRT models based on the normal ogive. Nearly 30 years, later he published the first volume that focused on IRT: *Applications of Item Response Theory to Practical Testing Problems* (Lord, 1980). With these two publications—and the numerous journal papers published between them—he may have done more than anyone else to advance the development of IRT.

Lord was a pragmatist who allowed the data to dictate the development of the model. In this vein, Lord was a developer and advocate of the two- and three-parameter IRT models.[23] In his work in the early 1950s, Lord coined several key terms, including ICC and test characteristic curve. He also described many foundational concepts of IRT, including local independence, item invariance, and standard errors conditional on test-taker proficiency (Carlson & von Davier, 2017; Lord, 1952). Lord continued to develop these concepts throughout his career.

Lord deserves considerable credit for development of the two- and three-parameter IRT models based on normal ogive, but it was Allan Birnbaum (1923–1976) who

reconceived Lord's work using the more mathematically tractable logistic models that are common in the early 21st century. In the late 1950s, Birnbaum was introduced to educational measurement when Columbia Teachers College invited him to help support a contract with the U.S. Air Force School of Aviation Medicine on test-taker classification (Barnard & Godambe, 1982). As we have already noted, in earlier work by Tucker, Green, and Lord, ICCs were described using a normal ogive model. Inspired by the work of David Haley (1952) and Joseph Berkson (van der Linden & Hambleton, 1997), Birnbaum believed the logistic form was more appropriate for IRT because of the existence of a sufficient statistic (Barnard & Godambe, 1982). To this end, in the late 1950s Birnbaum published several research reports for the U.S. Air Force on the applicability of the logistic model to IRT. Lord later invited Birnbaum to summarize this work in four chapters in *Statistical Theories of Mental Test Scores* (Birnbaum, 1968). In those chapters, Birnbaum provided the now-familiar logistic forms of the one-, two-, and three-parameter IRT models and introduced the scaling constant D to make the logistic form nearly identical to the earlier normal ogive formulation. Birnbaum also produced a practical method for item parameter estimation and later introduced the use of prior distributions into ability estimation (Birnbaum, 1967).

Birnbaum's contributions can reasonably be viewed as completing the foundations of IRT. As we have mentioned, related work also went on in Denmark. Whereas the models described in the previous paragraphs built on the insights of numerous researchers, the work in Denmark was carried out almost entirely by one man, Georg Rasch.

Development of the Rasch Model

Rasch (1901–1980) was born in Svendborg, Denmark. He studied mathematics at the University of Copenhagen and in 1930 earned his doctorate with the intention of pursing a teaching position at the university. When Rasch struggled to obtain a professorship in mathematics (during the worldwide depression), he turned his attention to statistical analysis and was able to earn a scholarship to study with the Nobel Prize–winning econometrician Ragnar Frisch and later with R. A. Fisher. He returned to Denmark in 1936 and was invited to teach statistics in the psychology department at the University of Copenhagen (Anderson & Olsen, 2001). It was during this time that he became increasingly interested in the problems of educational scaling. During the next 2 decades, Rasch worked as a statistical consultant, frequently on psychological and educational assessments. In 1952, while analyzing data for a Danish military exam, Rasch began work on the model for which he has become known.

Unlike Lord, who was a pragmatist, Rasch came to believe that it was critical to define the requirements necessary for objective measurement. By the time he published his monograph *Probabilistic Models for Some Intelligence and Attainment Tests* (Rasch, 1960/1980), he had developed the principles that have become associated with Rasch measurement: the requirements of sufficient statistics and specific objectivity. It is noteworthy that Rasch's *Probabilistic*

Models did not explicitly include the model widely associated with Rasch today—that is, the model equivalent to the one-parameter IRT model. The equivalent model presented in that volume lacked the logistic transformation. He began by defining the probability of a correct response as the ratio of the test takers' proficiency and the item difficulty.

> If we put $\xi/\delta = \zeta$, the problem is to choose a function of ζ which only takes values between 0 and 1. And as we find both very easy and very difficult problems and both very able and very unable persons, ζ must cover all values from zero to infinity.
>
> The simplest function I know of, which increases from 0 to 1 as ζ goes from 0 to ∞, is $\frac{\zeta}{1+\zeta}$. If we insert $\zeta = \frac{\xi}{\delta}$ we get $\frac{\frac{\xi}{\delta}}{1+\frac{\xi}{\delta}} = \frac{\xi}{\xi+\delta}$. (Rasch, 1960/1980; p. 74–75)

If we reparameterize by changing ξ to e^{θ} and δ to e^{b}, then the model becomes the typical Rasch model or the one-parameter logistic IRT model.

Several authors noted this relationship at approximately the same time. Birnbaum (1968, p. 402) noted that Rasch's model had this characteristic in the context of his more general presentation of the one-, two-, and three-parameter logistic models. Similarly, Wright (1968) included a footnote commenting that Rasch's model could be reformulated in logistic form.[24] It is unclear which of these authors has priority in identifying this relationship.[25]

After Rasch's (1960/1980) monograph, he went on to publish several additional papers on the model. Rasch's Berkeley Symposium paper from 1961 provides a general case for *specific objectivity*, and Rasch (1966) discussed other issues, including conditional independence. Throughout his work, Rasch discussed models that included logistic transformations, so it is a curiosity that he never introduced the now-common formulation of the logistic Rasch model for dichotomous items, $P(\theta) = \left(1+e^{\theta-b}\right)^{-1}$. Mark Reckase (personal communication, June 2, 2022) noted that it could be argued that Equation 4.2 in Rasch (1961) is the general case of all the models currently referred to as Rasch models and the simple logistic model is a special case of that equation. It may be that Rasch simply believed that it was more natural to use proficiency and difficulty scales with a lower bound of zero.

Although this reparameterized Rasch model for dichotomous items is equivalent to the one-parameter logistic IRT model, separate terms are often used to describe them to highlight the philosophical distinction between Rasch's approach and that represented by Lord's work.[26] Given the attention that has been placed on these philosophical distinctions, the reader might get the impression that Rasch developed his model based on principled a priori assumptions about the requirements for measurement; Rasch's own description of the development of his model does not support that interpretation. Regarding his early work on "reading capability," Rasch stated, "When I got the data I made the guess that it might be a good idea to try the multiplicative Poisson model. It turned out to fit the data quite well" (1960/1980, p. xiii). The con-

cept of *specific objectivity*, which has become a hallmark of Rasch measurement, seems to have been discovered after the fact, rather than motivating the development of the model. Later in life, Rasch made it clear that he only appreciated this characteristic of the model after discussing his work with his former mentor, Ragnar Frisch. Rasch visited Frisch in 1959 and described the work he had been doing in educational measurement. On seeing Rasch's formulation, Frisch expressed particular interest in the idea that the person and item parameters could be separated. This separability is what Rasch came to refer to as specific objectivity. Rasch went on to say, "What Frisch's astonishment had done was to point out to me that the possibility of separating two sets of parameters must be a fundamental property of a very important class of models" (1960/1980, p. xviii).

Rasch Versus Multiparameter IRT

This philosophical fault line that separated IRT and Rasch measurement became a growing fissure in the field of educational measurement from the late 1960s into the 1990s, often producing heated interactions at professional conferences. One such interaction was the 1992 debate between Benjamin Wright (1926–2015) and Ronald Hambleton (1943–2022) at the American Educational Research Association's annual meeting in San Francisco. To provide a sense of the division that existed in the field, we describe that debate in some detail.

Wright and Hambleton met in a packed ballroom to participate in an invited session titled "Which Models Work Best." Wright opened the debate: "I make measures for a living. Measures have a specific definition." He went on to suggest—wrongly—that the Rasch model was developed nearly a decade before multiparameter IRT. He proceeded to criticize multiparameter IRT for its "promiscuity" and willingness to "swallow whatever junk happens to come [its] way." Throughout his remarks, he repeatedly returned to the metaphor of a ruler that fails to provide objective measurement, saying, "no scientist, engineer, businessman or cook, who depends on measures . . . can work with that kind of ruler." He closed his remarks by attacking multiparameter IRT for being data centered, saying that the "Rasch model is theory-centered: data must fit, else get better data."

The younger Hambleton took a more genteel and structured approach in his response. He countered Wright's remarks on the antecedence of Rasch's work, noting that Lord's early work was published in 1952, well before Rasch's work was published in 1960. He went on to stress that the model must fit the data, saying that when "Rasch's work became known in 1960, . . . Lord had already rejected that model due to its inferior fit to his data." He discussed the virtues of modeling item discrimination and guessing, noting the 80-year history of modeling discrimination and the improvements to model fit that these parameters afforded. Hambleton devoted significant time to reflecting on the practicality of implementation, noting that the software is sufficient and improving and that many testing programs are having success with multiparameter IRT. Finally, he closed by calling for pragmatism and making it clear that there is a place for many different models, including Rasch's, saying, "Rasch has an important role only when it

fits data well, and sample sizes are modest with no need for highly precise estimates" (Hambleton, 1992; Huang, 2015; Wright, 1992).

Neither Wright nor Hambleton was declared the winner, but with the benefits of hindsight we can see that the winner of the larger debate within the field would likely not have satisfied either side. There can be little doubt that among organizations using IRT for operational purposes the one-parameter models are the most popular. So, in a sense, Wright's model was victorious. Importantly, however, Wright's philosophy was not. Operational testing organizations are not choosing a Rasch model because it reflects their philosophical view of objective measurement; instead, they are acceding to Hambleton's calls for pragmatism. The one-parameter model is popular because it accommodates modest sample sizes and is easy to implement for practitioners. In that way, it is safe to say that Hambleton's philosophy won out, even if his preferred model did not.

Although Wright and Hambleton famously disagreed about the appropriate philosophical lens through which to view these models, they shared quite a bit in common. Both men made substantial contributions to the popularization of IRT. Both men wrote articles and influential books on their vision of IRT, and both were influential professors—Wright at the University of Chicago and Hambleton at the University of Massachusetts Amherst—who trained a generation of psychometricians in the use of IRT.

IRT Extensions

Almost as soon as the foundations of IRT were completed, researchers began developing extensions to both Rasch and multiparameter IRT to accommodate a wide variety of response data. This process was so productive with extensions building on extensions that van der Linden and Hambleton's (1997) *Handbook of Modern Item Response Theory* presents no fewer than 27 models, including everything from polytomous extensions of the logistic models, to multidimensional extensions, to models incorporating response time.

One of the earliest IRT extensions is the result of work by Fumiko Samejima. Inspired by Lord's early work in 1952, Samejima became interested in IRT and began considering the possibility of polytomous extensions. After graduating in 1965, Samejima took several positions in North America, first as a visiting research psychologist at ETS, where she met Lord. After a number of other short-term university positions, in 1973 she accepted a position at the University of Tennessee, Knoxville. Samejima published work describing both normal ogive and logistic versions of what she called the *graded response model* (Samejima, 1969, 1972). Her approach was to model the probability that a test taker of a given proficiency would score at or above a given level on a polytomously scored task (Barnhart, 2013).

The next major extension to IRT models was the *nominal response model* developed by Darrell Bock (1927–2021) in 1972. After graduating from the University of Chicago, Bock worked at the University of North Carolina at Chapel Hill, where he was briefly a colleague of Fumiko Samejima, before returning to teach at the University of Chicago in the same department as Ben Wright. Bock's nominal response model was

conceptually similar to Samejima's heterogeneous graded response model but allowed for unordered categories. As a result, it was useful for the sort of unordered data that result from some psychological assessments as well as analyzing distractor patterns within educational assessments (Wainer & Robinson, 2006). David Thissen and Lynne Steinberg (1984) extended this latter use with their multiple-choice model. The multiple-choice model extends the nominal response model to support a guessing parameter to further support distractor analysis in education assessments.

The first polytomous extension from the Rasch perspective was provided by David Andrich (1978). After completing his undergraduate education in Australia, Andrich came to the United States to attend graduate school at the University of Chicago. Although Andrich worked with both Ben Wright and Darrell Bock, there is no doubt that he approached his work from a Rasch perspective. In fact, in 1977, Andrich spent 6 months working directly with Georg Rasch at the Danish Institute for Educational Research. Andrich's *rating scale model* estimates the probability that a person with a given ability would achieve each of the possible score points (Webster, 1998). The model, in addition to assuming that the slopes of the items are fixed, assumes that the threshold structure is consistent across all items on the exam. That is, the offsets between each score point and the nominal item difficulty are consistent across all items. Although this assumption may be difficult to satisfy for educational assessment, Andrich was primarily interested in Likert-type data, where the assumption may be more likely to be satisfied. In 1982, Geoff Masters, a fellow Australian and graduate of the University of Chicago, extended the rating scale model to make it more appropriate for educational assessments with the partial credit model. The partial credit model allowed the threshold structure of polytomous items to vary across the exam. This makes it possible for the difference between two specific score points to vary by item. As the name suggests, this can be applied in an educational context to provide some credit for a response that demonstrates partial mastery.

In 1992, the partial credit model was extended again by Eiji Muraki. Muraki, like Andrich and Masters, attended the University of Chicago for graduate school. Unlike Andrich and Masters, Muraki was advised by Darrell Bock and brought a multiparameter perspective to his extension of the partial credit model. In 1992, while studying at Chicago, Muraki published the *generalized partial credit model* that allowed the slope of the items to vary across the exam (Muraki, 1992). At roughly the same time, Wendy Yen, while working at CTB, independently developed the same model (Yen & Fitzpatrick, 2006). Although it was parameterized differently, Yen's two-parameter partial credit model and the generalized partial credit model are mathematically equivalent. This work produced independently by Muraki and Yen transformed the Rasch derived partial credit model to make it compatible with assessments that use a multiparameter IRT model.

Computerized Adaptive Testing

One prominent application of IRT is in the area of multistage and computerized adaptive testing, or what Lord originally referred to as "tailored testing." Broadly speaking, these approaches route test takers to different items based on proficiency estimates that are produced iteratively as the test is administered. These approaches had been explored in a classical test theory context in the 1950s and 1960s with Angoff and Huddleston's (1958) work on two-stage testing and Cleary et al.'s (1968) work on a variety of "programmed" testing methods, but these ideas needed an IRT framework to be practical. By the late 1960s, Lord had built on his earlier work on IRT to lay the theoretical foundation for adaptive testing and developed IRT-based methods of delivering both multistage tests with a router test and item-level computerized adaptive tests (Carlson & von Davier, 2017; Lord, 1968). By the early 1970s, Lord had gone further to address many of the practical challenges of adaptive testing, before the technology to deliver these tests was broadly available (Lord, 1970, 1971a, 1971b, 1971c, 1971d; Wainer, 2000).

Since the earliest conceptualizations of computerized adaptive testing there has been ongoing research on the best method for selecting items to be administered to each test taker in a manner that maximizes score precision while respecting nonstatistical constraints on item selection. One of the most popular solutions to this problem known as the *weighted deviation method* was introduced by Lord's longtime collaborator, Martha Stocking (1942–2006). Lord hired Stocking in 1967 as a research scientist at ETS after she completed her master's degree in statistics at Rutgers University (Bennett & von Davier, 2017; "Martha L. Stocking Swanson," 2006). Although her most famous and broadly applied contribution is undoubtedly the *Stocking–Lord equating method* (Stocking & Lord, 1983), Stocking was also prolific in the area of computerized adaptive testing, where she published several important articles and reports on practical considerations in controlling items exposure within an adaptive framework (Stocking et al., 1991a, 1991b; Stocking & Lewis, 1998, 2000; Stocking & Swanson, 1993, 1998).

An alternative approach to selecting items within a computerized adaptive test, while respecting content constraints, was proposed by Howard Wainer in 1987 (Wainer & Kiely, 1987). Wainer was hired in 1980 as a principal research scientist at ETS, where he remained for 21 years (Robinson, 2005). While there, as a colleague of both Lord and Stocking, he developed the testlet-based approach to selecting items within an adaptive test. In the testlet-based approach, rather than selecting individual items, small groups of items, or *testlets*, are administered as a set. These predefined groups of items can be selected and reviewed by content experts to ensure they meet all blueprint constraints before the exam is administered. Although this provides greater control over test content, it provides less adaptability than approaches that select one item at a time (van der Linden, 2000).

A third approach to item selection within computerized adaptive tests was developed by Wim van der Liden (2000). This approach, known as a *shadow test*, attempted to

allow for selecting items one at a time, while ensuring that the complete form adheres to all the underlying content constraints. In a shadow test, after each item is administered, the selection algorithm selects an entire form for that test taker that includes all previously administered items and meets all blueprint constraints. The selection algorithm then selects the item from this shadow test that provides the most information given the test taker's current ability estimate. Given an appropriate item pool, this approach ensures that the test taker will receive an exam that meets the blueprint constraints and is tailored to their ability.

The foundational theory, model extensions, and practical developments including improved estimation procedures (see Bock, 1997; Luecht & Hambleton, 2021; and Thissen & Steinberg, 2020) provided a basis that has allowed IRT to become largely ubiquitous in educational measurement. The third and fourth editions of *Educational Measurement* included major chapters on the topic. The applications are now so diverse and widespread that the current edition recognizes that the topic can no longer be covered in a single chapter.

In the prior three sections we focused on theoretical and technical developments that have shaped the current practice of educational measurement. In the next section we return to a focus on the practice of assessment.

LARGE-SCALE TESTING

In the second section, we described how intelligence testing grew from its foundations in the work of Galton, Spearman, and the eugenics movement. Intelligence testing represents a critical part of the history of the practice of testing, but it is only one part of that history. There is an even longer history of standardized tests being used for selection, graduation, and the evaluation of individuals and schools. Over time, these efforts developed into large-scale testing programs. Much of this testing is targeted at achievement rather than intelligence.

The Chinese began using written examinations for civil service selection more than 3,000 years ago when candidates were tested in the "six arts": music, archery, horsemanship, writing, arithmetic, and ceremonies of public life. Two thousand years ago, written tests were introduced in the "five studies": civil law, military affairs, agriculture, revenue, and geography of the empire. Later tests emphasized remembering and interpreting the writings of Confucius. Even for these early examinations there was an understanding that standardization was critical (DuBois, 1970).

The practice—and history—of Chinese testing was known in England and influenced testing practice there. In 1833, standardized testing was instituted for the British Indian civil service. Over time, similar practices were adopted in the United States. In the 1860s, a bill was presented in Congress to institute a similar civil service system, but that bill was never enacted. Civil service testing was eventually adopted in the United States, but that adoption occurred in a piecemeal fashion. The various efforts impacted specific departments in the government and came and went over time as funding was

approved or ran out. A permanent civil service commission was finally established in 1883.

Although university-based tests can hardly be viewed as large-scale assessments, to provide context we note that examinations have played an important role in that context as well. Although oral examinations may have been part of Western education since the founding of the first universities, written tests also have a long tradition. The written examination that Karl Pearson completed to achieve the rank of third wrangler at Cambridge, the Mathematical Tripos, dates to the 18th century. By the late 19th century, written examinations were common in universities.

Horace Mann and Testing in Public Schools

Given the price of paper, it is not surprising that the adoption of written testing in grammar schools lagged behind that in universities. In the United States, grammar school students were typically assessed based on oral recitation in the first half of the 19th century. The drive to create standardized written student assessment in the United States was motivated by the desire to move beyond judgments about students to evaluation of the performance of schools. By early-21st-century standards, the suggestion that standardized testing within a single city should qualify as *large-scale testing* may seem quaint, but nonetheless large-scale testing in the United States appears to have its foundations in the tests administered in the Boston public schools at the instigation of Horace Mann (1798–1859).

Mann is an icon of American education; he fought for universal nonsectarian education, and in recognition of that effort, schools across the country are named in his honor. He had essentially no experience as a teacher and yet he had exceptionally strong opinions about teaching. Most important in the context of this chapter, Mann saw that written tests were not only useful to evaluate individuals; the tests could also be used to evaluate schools—and, by extension, school administrators.

Prior to the launch of these tests in 1845, Boston's schools were evaluated by local visiting committees. The individual schools were run by masters with little additional oversight. The visiting committees routinely found the schools to be performing well based on idiosyncratic and subjective criteria. Understandably, the masters wished to be left alone to run their schools as they thought best (Reese, 2013). By contrast, Mann and his allies believed standardized testing would bring a measure of quality control to a bureaucratically managed and centralized education system. They also believed that through this mechanism they could establish performance standards and systematize education, teacher evaluation, and pedagogy.

In late spring 1845, after months of controversy between Mann and the Boston masters, the first large-scale written examinations were administered to students in Boston. As Mann and his allies planned, the examination results were used both to understand student learning and to challenge the authority of the masters. School accountability testing arguably grew from Boston's 1845 examinations, and Mann's reform work in Boston started a trend toward standardized testing throughout the United States,

including most notably the New York Regents Examinations. The early Regents examinations were introduced in 1865 following the model of Boston's written examinations. The initial examinations were used for high school admissions and to provide an annual measure of student achievement (Phelps, 2007).

The Regents examinations were the vanguard of competitive testing in U.S. schools. As school enrollment increased through the late 1800s, testing expanded beyond classroom assessments to include standardized assessments for grade promotion, high school admission, and high school graduation. Leading educators of the time supported increasingly consolidated schools and standardized curricula as well as standardized examinations, which were thought to spur excellence in teachers and students while preparing students for challenges outside school settings (Reese, 2013).

The written examinations introduced by Mann and the New York Regents established the use of standardized written examinations for both selection and accountability in the United States. Intelligence testing then went on to influence standardized selection and accountability tests after World War I. This progression and interaction will be apparent when we discuss the evolution of the SAT. However, there is an independent strand in the history of testing that warrants attention; that strand represents the continued use of tests to assess achievement, both for accountability and to evaluate the progress of individual students. J. M. Rice is a noteworthy part of that strand.

Joseph Mayer Rice

Joseph Mayer Rice (1857–1934) remains relatively unknown, although his work laid the groundwork for research by better known leaders in educational research, including E. L. Thorndike and Terman (Graham, 1966). He trained and practiced as a pediatrician and, at some point, became interested in pedagogy. This interest motivated his travels in Europe to learn about the developing field of psychology and contemporary theories of pedagogy. After returning to the United States, he developed and implemented tests of spelling, mathematics, and language to evaluate the impact of the pedagogical methods used across schools in cities around the country. He published the results from these studies in the popular magazine *The Forum* (Rice, 1897a, 1897b); many of his original articles were included in his best-known work, *Scientific Management in Education* (Rice, 1913).[27]

Rice succeeded in collecting large data sets to support his critique of the less-than-progressive educational practices that were prevalent at the time. His first study focused on spelling. In describing this work, Rice reported, "During this time three different tests were made; the number of children examined reaching nearly 33,000" (1897a, p. 164). The primary motivation of this data collection was to explore the relation between the amount of time used to drill spelling in the classroom and the outcomes measured by Rice's tests. Rice's contemporary, E. L. Thorndike, criticized Rice for being too quick to jump to speculative conclusions in interpreting his results (E. L. Thorndike, 1903). Nonetheless, it is reasonable to consider this data collection to be the beginning of the science of educational research. Stanley (1966) suggested that this was likely "the first

full-scale 'experiment' ever done in schools and published" (p. 135). It is similarly not difficult to find evidence to support the conclusion that Rice was a self-promoter who brought muckraking to the field of education (Graham, 1966).

School-Based Achievement Testing

In a previous section we discussed the rise of intelligence testing following World War I. At that same time, the use of achievement tests also grew dramatically. Students were often tested with both intelligence and achievement tests in an effort to evaluate whether their performance was in line with their potential.[28] The other SAT—the Stanford Achievement Test—was developed by Kelley, Ruch, and Terman in 1922. An advertisement for the test in Kelley (1927, end of volume) somewhat ironically describes the test as follows:

> This new battery of achievement tests is designed to measure very thoroughly the knowledge and ability of pupils in school subjects in grades two through eight. It covers all the ground necessary to cover for ordinary purposes of educational testing.
>
> The score in any subject is immediately comparable with the score in any other subject, and valid composite scores for any number of subjects taken together are readily obtainable. Age norms as well as grade norms make possible the derivation of a satisfactory educational quotient.

The battery included tests of arithmetic, reading, spelling, science information, and history and literature. A century after its introduction, the test—now in its 10th revision—is still in use.

In 1927, the Stanford Achievement Test was far from the only large-scale test available. Kelley (1927) listed 4 elementary school achievement batteries, as well as 12 elementary intelligence tests, 15 elementary reading tests, and 10 elementary spelling tests. Clearly, the decade after the First World War was a period of substantial growth for standardized testing.

Over time, other achievement tests were introduced. Perhaps the best known was the Iowa Test of Basic Skills, which continues to the present day as part of the Iowa Assessments. These tests supported a substantial testing industry, including for-profit corporations such as CTB/McGraw–Hill, Riverside, and the World Book Company. These large-scale achievement tests also provided motivation for improving testing procedures. Perhaps most notably, E. F. Lindquist—who was a major force behind the Iowa testing program—introduced the first optical scanner to score multiple-choice items.

So far, this section has focused on achievement tests. These tests were intended as accountability measures for schools, to assess individual progress, and as a means of evaluating the efficiency of differing pedagogical approaches. We now consider two tests used for college admissions: One had its origins in intelligence testing and the other in achievement testing.

The SAT

The group that developed and administered the Army Alpha test included luminaries of the science of intelligence testing such as Yerkes, Terman, Goddard, and Otis. The group also included Carl Brigham (1890–1943). The Army Alpha test was used to measure the "verbal ability, numerical ability, ability to follow directions, and knowledge of information" of military recruits. As we have seen, Army Alpha helped to launch the now-ubiquitous multiple-choice question; it also introduced intelligence testing on a large scale. Perhaps less well known is the fact that it was a steppingstone to the creation of the Scholastic Aptitude Test, the precursor to the current SAT. The connection between the two tests is straightforward; after the war, Brigham adapted the Army Alpha test to create the original SAT.

Established in 1900, the College Entrance Examination Board represented member colleges in the interest of developing uniform assessments (essays at the time) aligned with the curricula of schools that supplied graduates to elite institutions of higher education (Lemann, 2004). Previously, individual schools were responsible for developing their own admissions essays, and the College Board examinations introduced uniformity into the administration process, the tested content, and the scoring. The College Board essay tests were first administered in 1901 (Fuess, 1950). Though the goal of these examinations was standardization, College Board member schools expressed concerns in the 1920s about the subjectivity in scoring and the narrowing effect the exams had on high school curricula (Bennett, 2017).

These concerns led to the development of alternative assessment approaches, including multiple-choice achievement and aptitude tests, leading to the development and adoption of Brigham's Scholastic Aptitude Test[29] in 1926 (Lemann, 2004).

Led by James Bryant Conant, then-president of Harvard University, in collaboration with Brigham and Henry Chauncey (then–assistant dean of admissions, Harvard University), the test was first employed to identify scholarship students based on their intellect, not their educational background. Harvard began using the test in 1934 to identify scholarship students, eventually requiring it of all applicants beginning in 1941. It became a requirement of College Board member schools in 1942.

The first version of the SAT, in use from 1926 to 1930, comprised nine subtests containing 315 multiple-choice questions; it was administered under highly speeded conditions, allowing 97 minutes for the full administration. The more familiar two-section format, testing verbal and mathematical aptitude, was introduced in 1930 (Lawrence et al., 2004).

As discussed earlier, intelligence testing rose to prominence during this same period. By 1932, a vast majority of large school systems in the United States were using intelligence testing to guide ability grouping (Haney, 1984). Despite the similarities in general appearance between intelligence tests and contemporaneous versions of the SAT, there is a marked tension between intelligence testing and the SAT. Throughout much of its existence, the developers of the SAT have argued that instead of measuring intelligence, versions of the SAT have been designed to measure aptitude and reasoning.

Currently, the College Board claims that its questions focus on skills that matter for college success.

ACT (Formerly American College Testing)

Although the SAT was the first large-scale college admissions test, it is no longer the largest. In response to the increasing number of college applicants following World War II and the federal initiatives of the Great Society to increase access to higher education, there was a high demand for standardized tools to support college admissions decisions. The SAT was commonly used in the northeastern states, but elsewhere in the country, states developed their own admissions tools, which increased the complexity of college application, particularly for out-of-state students (ACT, 2009). The American College Testing Program, also known as the ACT Test and ACT,[30] was an attempt to meet this need. The program had its roots in the Iowa Academic Meet, the first Iowa testing program for high school students.

The Iowa Academic Meet, or the "Brain Derby," was developed in 1929 by E. F. Lindquist (editor of the first edition of *Educational Measurement*) under the direction of Professor Thomas Kirby (Holmgren, 2009). The Iowa Meet, in its early form, was used to identify outstanding scholars while raising the standards of instruction across Iowa high schools by testing every student (Lindquist, 1976, as referenced by Croft & Beard, 2021). This grew into the Iowa Assessments (previously known as the Iowa Testing Programs). Iowa Assessments were administered to students in Iowa beginning in 1935 and around the country beginning in the 1940s.

The ACT, cofounded by Lindquist and Ted McCarrell of the University of Iowa, was introduced in 1959 and was designed to assess a student's general educational development. It incorporated multiple-choice questions that had previously appeared on the Iowa Test of Educational Development, allowing for rapid development (Croft & Beard, 2021). The original ACT had four sections—English, mathematics, social studies, and science—and was delivered in 3 hours (Croft & Beard, 2021; see ACT, 2009, for further details).

The ACT was originally tied to the assessment of knowledge, in contrast to the SAT's focus on aptitude. Although this historical distinction exists, it would appear to be of little practical importance given the high correlation that exists between the tests—typically reported to be approximately .90—and the existence of widely used concordance tables that allow the scores to be used interchangeably (College Board & ACT, 2018; Perez, 2002).

In the previous pages, we considered tests designed for accountability, tests designed as a metric of individual student achievement, and tests for selection. So far, the tests we have discussed have one thing in common: They produce an interpretable score for each test taker. We conclude this section by considering a test explicitly designed so that individual scores cannot be produced.

NAEP

In 1867, Congress established a Department of Education. Among the purposes of the department was "collecting such statistics and facts as shall show the condition and progress of education in the several States and Territories" (An Act to Establish a Department of Education, 1867, Sec. 1). In essence, the department was to produce a national assessment of educational progress. This did not happen; the Department of Education was reduced to an office and then made part of the Department of the Interior.

Nearly 100 years later, the idea of a national assessment of educational progress was revived. (See also Ho & Polikoff, this volume, for a review of the history and current use of NAEP for accountability purposes.) In 1964, the test was funded by the Carnegie Corporation. The Education Commission of the States subsequently took over administration with funding from the Carnegie Corporation and then with increasing funding from the U.S. Office of Education. Congress took over all funding for NAEP in 1972 and moved NAEP to the National Institute of Education in 1978.

From the outset there was an effort to make NAEP an outstanding example of measurement science. The office of education—and later the Department of Education—contracted and consulted with the most respected organizations in the country, including ETS, the American Council on Education, American College Testing (ACT), Human Resources Research Organization, and the American Institutes for Research. This led to innovation and the application of state-of-the-art methodology.

Ralph Tyler (1902–1994), a contributor to the first edition of *Educational Measurement*, was an early advisor on the project. It was Tyler's recommendation to base the national assessment on a systematic sampling strategy rather than a more exhaustive program of testing. Sampling reduced cost, but also ruled out the possibility of producing scores for individuals, which was prohibited by the legislation.

In 1983, the sophistication of the NAEP administration again took a major step forward. ETS was awarded a contract by the Department of Education to design and conduct the examination. ETS proposed numerous changes (Messick et al., 1983). They included a more sophisticated sampling design as well as the introduction of IRT to support test construction and to carry out linking across test forms and across years.

As we noted previously, NAEP has been an example of excellence and innovation in assessment science since the 1960s. In the IRT chapter from the fourth edition of *Educational Measurement*, Yen and Fitzpatrick (2006) described NAEP as "the most famous example of a matrix-sampled test." They go on to say that "NAEP is also notable for the use of IRT in combination with advances in missing data technology and hierarchical analyses to estimate population characteristics without estimating individual examinee scores" (p. 145). NAEP was referenced in five separate chapters in the 1989 edition of *Educational Measurement* and in eight separate chapters of the fourth edition (Brennan, 2006).[31]

As this section suggests, much of the development of large-scale testing has happened in the United States, but in recent decades these tests have become important around the globe. For example, in China, the National College Entrance Examination, which determines eligibility for postsecondary education, is taken by approximately 10 million students each year. Because of the impact this test can have on future career opportunities, students spend years preparing for it (Larmer, 2014). High school graduation and university selection tests are also an established part of education in many other countries (see Clauser & Margolis, 2023). Beginning in the 1990s, three large-scale achievement tests have also been administered internationally in multiple languages: these include the Trends in International Mathematics and Science Study (TIMSS, first administered in 1996), the Programme for International Student Assessment (PISA, first administered in 2000), and the Progress in International Reading Literacy Study (PIRLS, first administered in 2001). These international tests are discussed in detail in Chapter 20 of this volume by Braun and Kirsch.

THE HISTORY OF EDUCATIONAL MEASUREMENT AS REFLECTED IN FIVE EDITIONS OF *EDUCATIONAL MEASUREMENT*

The multiple editions of *Educational Measurement* now span more than 7 decades (Brennan, 2006; Lindquist, 1951; Linn, 1989; R. L. Thorndike, 1971), and in this final section we consider how these volumes fit into the history of *Educational Measurement*. Since the first edition was published, it has been a highly regarded standard reference. The volumes can fairly be seen as representing the state of the art for educational measurement, but in general they do not represent the cutting edge (for example, the 1971 edition included only passing mention of the relatively new technology represented by IRT, although it had been thoroughly described by Birnbaum in Lord and Novick's 1968 text and by Rasch in 1960).

Although there are lead authors responsible for the chapters in each of the volumes, the works have been intentionally collaborative. The Lindquist volume lists collaborators along with the lead authors on the title page of each chapter. Subsequent volumes all highlight the reviewers for each chapter. Although there is sometimes disagreement within the field about some of the views expressed, the resulting chapters are at least intended to make a consensus statement about their specific topics. As such, changes across chapters from one edition to the next provide a useful view of how the field has evolved. In reviewing the third edition, Cronbach (1989b) commented on the increased length of each subsequent volume, the topics covered, and the level of mathematical sophistication required of the reader. In addition to these changes across editions, we also consider how the authors have changed over 70 years.

A synoptic view of the four earlier volumes provides considerable insight. Each of the volumes has a section on measurement theory. In the third and fourth editions, this is broadened to include "theory and principles." In the first two editions, this section is

placed after the more practical sections on test development and administration. In the subsequent editions, theory is (literally) given priority by being placed at the beginning of the volume. This seems to reflect a change in the intended audience of the volumes; it may also reflect a change in the field, with higher expectations that practitioners have advanced training. In a sense, these subsequent volumes demonstrate that the place of the measurement profession was now established.

One interesting way to look across the editions is to consider how the content has changed in subsequent chapters on the same topic. Examining the parallel chapters provides a detailed outline of historical changes in thinking within the field. For example, the chapters on validity allow the reader to see how the concept grew from validity as correlation and prediction (in Cureton's 1951 chapter) to include construct validity (as described by Cronbach, 1971). Messick (1989) then broadened the view of construct validity and gave particular attention to consequences and Kane (2006) formalized the idea of validity as a structured argument in support of the inferences we make based on test scores. Green (2008) pointed out that in the fourth edition correlation and prediction are almost absent from the presentation of validity.

Kolen (2021), in discussing the history of equating, made a similar case for the historical value of the early editions of *Educational Measurement.* He described Flanagan's (1951) chapter as the first comprehensive description of equating and discussed how Angoff (1971)—writing in the second edition—reconsidered some of the basic assumptions of equating (e.g., the idea that equating should be sample invariant). As is the case with validity theory, changes in practice continue to be documented in the subsequent editions as equal percentile and IRT-based approaches receive more attention.

In addition to considering how similar content has evolved across parallel chapters in the editions of *Educational Measurement*, it is interesting to note how the content has changed across volumes. In all five editions the section on theory includes chapters on validity, reliability, and equating. The first two editions also included chapters related to theory, which were dropped from subsequent editions. Both the first and the second editions included chapters on approaches for considering multiple measurements (i.e., batteries and profiles). As less attention has been given to intelligence testing, these sections appear to have been considered less relevant. These two early editions also included chapters on the *nature of measurement*; this topic was not included in the third or fourth editions but has a reprise in the current edition.

Not surprisingly, the section on theory and principles in subsequent editions also includes chapters that were not included in the first two volumes. Three chapters in the Linn (1989) edition clearly reflect trends in the field that have received considerable attention since the second edition was published: IRT, cognitive psychology, and bias in testing. As Cronbach (1989a) pointed out, although by 1971 there had already been a long presentation of IRT by Birnbaum and a book on computer-assisted testing, there is almost no mention of IRT in the second edition. The attention given to this area since the 1970s is reflected by the chapters included in the third and fourth editions.

Similarly, the controversy over racial differences in IQ scores, concern for equity and fairness, and the increased attention given to consequences as an integral part of test validity led to a very significant focus on test bias. This is reflected in the chapter on bias in the third edition. Finally, chapters on cognitive psychology in the third and fourth editions reflect the extent to which measurement experts searched for a theory-based foundation for assessment practice.

In addition to considering the content of the editions, it is interesting to consider how the contributors have changed over time. In the first three editions, the majority of the authors were affiliated with universities (57%, 61%, and 69%, respectively). In the two most recent editions, fewer than 50% of the authors had university affiliations; most were associated with nonprofit organizations. Relatively few of the contributors to any of the editions have come from for-profit testing companies.

The authors who have made contributions to *Educational Measurement* have also changed in another important way over time. The list of authors in the first two editions included only one woman.[32] The percentage of women contributing to the volumes has substantially increased with the subsequent editions. Across the four earlier editions, the percentage of women authors was 4, 0, 27, and 35, respectively. In the current edition more than 40% of the authors are women.

SOME FINAL THOUGHTS

In this chapter we have described some of the critical themes and events in the history of educational measurement. We believe that understanding the history of our field is important. To use a phrase that Howard Wainer (2005, p. 118) attributed to Aristotle, "We understand best those things we see grow from their very beginnings." If we wish to understand the field of educational measurement, we would be well served by understanding how it developed. That history tells us something of what has been accomplished since the time of Francis Galton. It is easy to view the development of correlation by Galton and Pearson; the adoption of those methods for correlational psychology by Spearman; the extension of Spearman's formulas by Kelley, Kuder and Richardson, and Cronbach; and finally the development of generalizability theory as an inevitable evolution. The history as presented in this chapter documents that evolution, but it also makes clear that there was nothing inevitable about it. The dominant approach to understanding reliability could instead have been based on Edgeworth's generalizability-like framework—which was published before Galton's introduction of the correlation coefficient. The evolution of test theory, like Darwinian evolution, has not been a progression toward some abstract perfection; it simply reflects the relative usefulness of different approaches given the environment in which they exist. Spearman's work supported an exploration of the nature of intelligence, and that exploration was well suited to an environment shaped by the widely popular eugenics movement.

These beginnings have had a lasting impact on both the theory and the practice of testing. As we discussed in the second section, the Army Alpha test brought to prominence large-scale intelligence testing using multiple-choice questions. The focus on intelligence testing had a major impact on testing practice—up to and including early versions of the SAT, which were modeled after the Army Alpha test; the general format represented by these tests (if not the content) has remained dominant for over a century.

The early prevalence of intelligence testing also created a close link between testing theory and practice and the eugenics movement. This link helps to explain why the public may not trust experts who suggest that they can benefit society by measuring minds. The eugenics movement caused horrific suffering both in Europe and in the United States. Many of the great thinkers in statistics and educational measurement were closely aligned with that movement. Names like Galton, Pearson, Spearman, Fisher, Goddard, and Kelley represent a very partial list of eugenicists associated with educational measurement.

In understanding the context that has produced the practice of testing in the 21st century, it is also important to remember the political forces that have shaped testing. At least as far back as the 1860s, the U.S. Congress has made political decisions about whether to support—or to withhold support from—testing. The civil service commission and the army testing program during World War I are early examples, but the testing mandates put in place to evaluate the American education system and to provide accountability (such as NAEP) have likely done even more to shape the field of educational measurement.

Another example of the importance of context arises in the sections on scaling and the development of IRT. These sections are closely linked and both remind us that beliefs about the nature of educational measurement have not only changed over time, but also varied across researchers at the same time. Fechner constructed a scale that reflected his understanding of measurement from the study of physics. Binet, by contrast, rejected the link between psychological and physical measurement, stating, "The scale properly speaking does not permit the measure of intelligence, because intellectual qualities are not super-posable, and therefore cannot be measured as linear surfaces" (Binet & Simon, 1916, p. 40). This split between pragmatism and philosophical beliefs about the nature of measurement resulted in high-profile disagreements between the adherents of multiparameter IRT and Rasch measurement. That debate continues in more recent works such as those by Mari et al. (2021) and Irribarra (2021). Understanding how these differences of opinion have persisted should help the reader to hold these disagreements in perspective.

Each of the sections in this chapter provides an example of why we believe that the theory and practice of educational measurement needs to be understood in their historical context. It is easy to read a text on classical test theory or IRT and come away with the impression that they are inevitable truths describing immutable relationships. These theories are simply tools that have proven useful in specific contexts. These tools are valuable (and the present authors have spent decades trying to mas-

ter their use), but as the questions we ask change, the tools may need to change as well. Historical context provides an understanding of how and when those changes become critical.

In the end, it is not our intention to tell the reader how to interpret the history of our field. We certainly do not have answers to the question of what lessons should be learned from that history. But we do believe that leaders in educational measurement will understand the field better when they understand how it has grown from its beginnings.

ACKNOWLEDGMENTS

One of the great challenges of writing this chapter is the great expanse represented by the history of educational measurement. It would be impossible for any set of authors to know it all. We are, therefore, humbled by the tremendous support we have received throughout the process. First, we would like to thank Linda Cook and Mary Pitoniak for entrusting us with this work, as well as for their leadership and thoughtful suggestions throughout. We would also like to thank Howard Wainer, Mark Reckase, and Kadriye Ercikan, who acted as reviewers and collaborators on the chapter, for their multiple reviews filled with many valuable suggestions. Sometimes these suggestions corrected errors, but more importantly, they led us to include historical figures and events that we may otherwise have overlooked. Their contributions have improved both the content and the form of the chapter.

REFERENCES

Abelson, A. R. (1911). The measurement of mental ability of "backward" children. *British Journal of Psychology, 4*(3–4), 268–314. https://doi.org/10.1111/j.2044-8295.1911.tb00047.x

ACT. (2009). *ACT: The first fifty years, 1959–2009*.

An Act to Establish a Department of Education, Pub. L. No. 39-73, 14 Stat. 434 (1867). https://www.docsteach.org/documents/document/act-of-march-2-1867-public-law-3973-14-stat-434-which-established-the-department-of-education

Aikens, H. A., Thorndike, E. L., & Hubbell, E. (1902). Correlations among perceptive and associative processes. *Psychological Review, 9*(4), 374–382. https://doi.org/10.1037/h0070072

Andersen, E. B., & Olsen, L. W. (2001). The life of Georg Rasch as a mathematician and as a statistician. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 3–24). Springer. https://doi.org/10.1007/978-1-4613-0169-1_1

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573. https://doi.org/10.1007/BF02293814

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). American Council on Education.

Angoff, W. H., & Huddleston, E. M. (1958). *The multi-level experiment: A study of a two-stage test system for the College Board Scholastic Aptitude Test* (Statistical Report SR-58–21). ETS.

APF Gold Medal Awards and Distinguished Teaching of Psychology Award. (1993). *American Psychologist, 48*(7), 717–725. https://doi.org/10.1037/h0090746

Bagley, W. C. (1901). On the correlation of mental and motor ability in school children. *The American Journal of Psychology, 12*(2), 193–205. https://doi.org/10.2307/1412533

Barnard, G. A., & Godambe, V. P. (1982). Memorial article: Allan Birnbaum 1923–1976. *The Annals of Statistics, 10*(4), 1033–1039. https://doi.org/10.1214/aos/1176345968

Barnhart, R. (2013). *Fumiko Samejima (b. 1930)*. Society for the Psychology of Women. https://www.apadivisions.org/division-35/about/heritage/fumiko-samejima-biography

Bennett, R. E. (2017). What does it mean to be a nonprofit educational measurement organization in the twenty-first century? In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 1–15). Springer. https://doi.org/10.1007/978-3-319-58689-2_1

Bennett, R. E., & von Davier, M. (Eds.). (2017). *Advancing human assessment: The methodological, psychological and policy contributions of ETS*. Springer. https://doi.org/10.1007/978-3-319-58689-2_1

Binet, A., & Fere, C. (1887). *Le magnetisme animal*. Alcan.

Binet, A., & Simon, T. (1916). *The development of intelligence in children (the Binet–Simon Scale)* (E. S. Kite, Trans). Williams and Williams Co. https://doi.org/10.1037/11069-000

Birnbaum, A. (1957). *Efficient design and use of tests of ability for various decision-making problems* (Report No. 58-16). USAF School of Aviation Medicine.

Birnbaum, A. (1958a). *Further considerations of efficiency in tests of a mental ability* (Report No. 17). USAF School of Aviation Medicine.

Birnbaum, A. (1958b). *On the estimation of mental ability* (Report No. 15). USAF School of Aviation Medicine.

Birnbaum, A. (1967). *Statistical theory for logistic mental test models with a prior distribution of ability* (ETS Research Bulletin No. RB-67-12). ETS. https://doi.org/10.1002/j.2333-8504.1967.tb00363.x

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison–Wesley.

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16*(4), 21–33. https://doi.org/10.1111/j.1745-3992.1997.tb00605.x

Boring, E. G. (1950). *A history of experimental psychology* (2nd ed.). Prentice Hall.

Brennan, R. L. (1983). *Elements of generalizability theory*. ACT.

Brennan, R. L. (2001a). *Generalizability theory*. Springer-Verlag. https://doi.org/10.1007/978-1-4757-3456-0

Brennan, R. L. (2001b). *Manual for mGENOVA*. University of Iowa, Iowa Testing Programs.

Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). American Council on Education; Praeger.

Brennan, R. L. (2021). Generalizability theory. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 206–231). Routledge. https://doi.org/10.4324/9780367815318-10

Briggs, D. C. (2021). A history of scaling and its relationship to measurement. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 263–291). Routledge. https://doi.org/10.4324/9780367815318-12

Briggs, D. C. (2022). *Historical and conceptual foundations of measurement in the human sciences: Credos and controversies*. Routledge. https://doi.org/10.1201/9780429275326

Brigham, C. C. (1923). *A study of American intelligence*. Princeton University Press.

Brigham, C. C. (1930). Intelligence tests of immigrant groups. *Psychological Review, 37*(2), 158–165. https://doi.org/10.1037/h0072570

Brooks, J. L. (1984). *Just before the origin: Alfred Russel Wallace's theory of evolution*. Columbia University Press.

Brown, W. (1910a). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Brown, W. (1910b). *The use of the theory of correlation in psychology*. Cambridge University Press.

Brown, W. (1911). *The essentials of mental measurement*. Cambridge University Press.

Bunch, M. B. (2021). The role of the federal government in shaping educational testing policy and practice. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement*: *Key advancements in theory, policy, and practice* (pp. 65–86). Routledge. https://doi.org/10.4324/9780367815318

Burt, C. (1936). The analysis of examination marks. In P. Hartog & E. C. Rhodes (Eds.), *The marks of examiners* (pp. 245–314). Macmillan.

Burt, C. (1955). Test reliability estimated by analysis of variance. *British Journal of Statistical Psychology, 8*(2), 103–118. https://doi.org/10.1111/j.2044-8317.1955.tb00325.x

Capshew, J. H. (1999). *Psychologists on the march: Science, practice, and professional identity in America, 1929–1969*. Cambridge University Press. https://doi.org/10.1017/CBO9780511572944

Carlson, J. E., & von Davier, M. (2017). Item response theory. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 133–178). Springer. https://doi.org/10.1007/978-3-319-58689-2_1

Cherry, K. (2023, September 28). *Edward Thorndike's contributions to psychology*. Very Well Mind. https://www.verywellmind.com/edward-thorndike-biography-1874-1949-2795525

Clauser, B. E. (2021). A history of classical test theory. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement*: *Key advancements in theory, policy, and practice* (pp. 157–180). Routledge. https://doi.org/10.4324/9780367815318-8

Clauser, B. E., & Margolis, M. J. (2023). Past, present, and future of educational measurement. In R. J. Tierney, F. Rizvi, & K. Erkican (Eds.), *International encyclopedia of education: Vol. 14. Quantitative research and educational measurement* (pp. 1–14). Elsevier. https://doi.org/10.1016/b978-0-12-818630-5.10001-6

Clark, C. L. (Ed.). (1976). *Proceedings of the first conference on computerized adaptive testing.* U.S. Government Printing Office.

Cleary, T. A., Linn, R. L., & Rock, D. A. (1968). An exploratory study of programmed tests. *Educational and Psychological Measurement, 28*(2), 345–360. https://doi.org/10.1177/001316446802800212

Clifford, G. J. (n.d.). *Edward L. Thorndike (1874–1949): The man and his career, a psychology for educators, education as specific habit formation*. Education Encyclopedia. https://education.stateuniversity.com/pages/2509/Thorndike-Edward-L-1874-1949.html

College Board & ACT. (2018). *Guide to the 2018 ACT/SAT concordance*. https://satsuite.collegeboard.org/media/pdf/guide-2018-act-sat-concordance.pdf

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics, 27*(4), 907–949. https://doi.org/10.1214/aoms/1177728067

Croft, M., & Beard, J. J. (2021). Development and evolution of the SAT and ACT. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement*: *Key advancements in theory, policy, and practice* (pp. 22–41). Routledge. https://doi.org/10.4324/9780367815318-2

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). American Council on Education.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist, 30*(1), 1–14. https://doi.org/10.1037/0003-066X.30.1.1

Cronbach, L. J. (1989a). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Lawrence Erlbaum.

Cronbach, L. J. (1989b). Review of *Educational Measurement*, third edition. *Educational Measurement: Issues and Practice, 8*(4), 22–25. https://doi.org/10.1111/j.1745-3992.1989.tb00339.x

Cronbach, L. J. (1991). Methodological studies—a personal perspective. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 385–400). Lawrence Erlbaum.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. John Wiley & Sons.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*(2), 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391–418. https://doi.org/10.1177/0013164404266386

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). American Council on Education.

Darwin, C. (1859). *On the origin of species*. John Murray.

Darwin, C. (1869, December 23). [Letter to Francis Galton]. Darwin Correspondence Project (Letter no. 7032). https://www.darwinproject.ac.uk/letter/DCP-LETT-7032.xml

Dorans, N. J. (2004). *A conversation with Ledyard R. Tucker*. ETS. http://www.ets.org/Media/Research/pdf/TUCKER.pdf

Dr. Paul Lazarsfeld dies. (1976, September 1). *The New York Times*. https://www.nytimes.com/1976/09/01/archives/dr-paul-lazarsfeld-dies-sociologist-at-columbia.html

DuBois, P. H. (1970). *A history of psychological testing*. Allyn & Bacon.

Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society, 51*, 346–368. https://doi.org/10.1177/014662168400800104

Education Amendments of 1974, Pub. L. 93-380, 88 Stat. 484 (1974). https://www.govinfo.gov/content/pkg/COMPS-727/pdf/COMPS-727.pdf

Engelhard, G., Jr. (1984). Thorndike, Thurstone, and Rasch: A comparison of their methods of scaling psychological and educational tests. *Applied Psychological Measurement, 8*(1), 21–38. https://doi.org/10.1177/014662168400800104

Fechner, G. T. (1966). *Elements of psychophysics* (H. E. Adler, Trans.; D. H. Howes & E. G. Boring, Eds.; Vol. 1). Rinehart & Winston. (Original work published 1860)

Fisher, I. (1923, December 21). Letter to Truman Kelley from the Eugenics Committee of the United States of America. Harvard University Archives: Truman Kelley Papers.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd.

Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd.

Flanagan, J. C. (1939). *The cooperative achievement tests. A bulletin reporting the basic principles and procedures used in the development of their system of scaled scores*. American Council on Education Cooperative Test Service.

Flanagan, J. C. (1942). A study of factors determining family size in a selected professional group. *Genetic Psychology Monographs, 25*, 3–99.

Flanagan, J. C. (Ed.). (1948). *The Aviation Psychology Program in the Army Air Forces* (Vol. 1). U.S. Government Printing Office. https://doi.org/10.1037/e614382011-001

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

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*(4), 327–358. https://doi.org/10.1037/h0061470

Fuess, C. M. (1950). *The College Board: Its first fifty years*. Columbia University Press.

Galton, F. (1853). *Tropical South Africa*. John Murray.

Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences*. Macmillan. https://doi.org/10.1037/13474-000

Galton, F. (1875). Statistics by intercomparison with remarks on the law of frequency of error. *Philosophical Magazine, 49*(322), 33–46. https://doi.org/10.1080/14786447508641172

Galton, F. (1884). *Anthropometric laboratory*. William Clowes and Sons.

Galton, F. (1886a). Family likeness in eye-colour. *Proceedings of the Royal Society, 40*(242–245), 402–417. https://doi.org/10.1098/rspl.1886.0058

Galton, F. (1886b). Family likeness in stature. *Proceedings of the Royal Society, 40*(242–245), 40–72. https://doi.org/10.1098/rspl.1886.0009

Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society, 45*(273–279), 135–145. https://doi.org/10.1098/rspl.1888.0082

Galton, F. (1889). *Natural inheritance*. Macmillan. https://doi.org/10.5962/bhl.title.32181

Galton, F. (1892). *Finger prints*. Macmillan.

Galton, F. (1893). *Decipherment of blurred finger prints*. Macmillan.

Galton, F. (1895). *Finger print directories*. Macmillan.

Galton, F. (1908). *Memories of my life*. Methuen. https://doi.org/10.5962/bhl.title.28398

Galton, F. (1909). *Essays in eugenics*. Eugenics Education Society. https://doi.org/10.5962/bhl.title.168760

Gillham, N. W. (2001). *The life of Sir Francis Galton: From African exploration to the birth of eugenics*. Oxford University Press. https://doi.org/10.1093/oso/9780195143652.001.0001

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika, 30*, 395–418. https://doi.org/10.1007/BF02289531

Goddard, H. H. (1912). *The Kallikak family: A study in the heredity of feeble-mindedness*. Macmillan.

Gottfredson, L. S. (1994, December 13). Mainstream science on intelligence. *Wall Street Journal*, A18.

Gottfredson, L. S. (2018). G theory: How recurring variation in human intelligence and the complexity of everyday tasks create social structure and the democratic dilemma. In R. J. Sternberg (Ed.), *The nature of human intelligence* (pp. 130–151). Cambridge University Press.

Gould, S. J. (1996). *The mismeasure of man*. Norton & Company.

Graham, P. A. (1966). Joseph Mayer Rice as a founder of the progressive education movement. *Journal of Educational Measurement, 3*(2), 129–133. https://doi.org/10.1111/j.1745-3984.1966.tb00868.x

Green, B. F. (1951a). *Latent class analysis: A general solution and an empirical evaluation* (ETS Research Bulletin No. RB-51-15). ETS. https://doi.org/10.1002/j.2333-8504.1951.tb00215.x

Green, B. F. (1951b). A general solution for the latent class model of latent structure analysis. *Psychometrika, 16*, 151–166. https://doi.org/10.1007/BF02289112

Green, B. F. (2008). Book review: *Educational measurement* (4th ed.). *Journal of Educational Measurement, 45*(2), 195–200. https://doi.org/10.1111/j.1745-3984.2008.00060.x

Gulliksen, H. (1936). The content reliability of a test. *Psychometrika, 1*, 186–194.

Gulliksen, H. (1950). *Theory of mental tests*. Wiley. https://doi.org/10.1037/13240-000

Haley, D. C. (1952). *Estimation of dosage mortality relationship when the dose is subject to error* (Technical Report No. 15). Applied Mathematics and Statistics Laboratory, Stanford University.

Hambleton, R. K. (1992). Hambleton's 9 theses. *Rasch Measurement Transactions, 6*(2), 215–217. https://www.rasch.org/rmt/rmt62d.htm

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer–Nijhoff Publishing. https://doi.org/10.1007/978-94-017-1988-9

Haney, W. (1984). Testing reasoning and reasoning about testing. *Review of Educational Research, 54*(4), 597–654. https://doi.org/10.3102/00346543054004597

Hart, B., & Spearman, C. (1912). General ability, its existence and nature. *British Journal of Psychology, 5*(1), 51–84. https://doi.org/10.1111/j.2044-8295.1912.tb00055.x

Herrnstein, R. J. (1971). I.Q. *Atlantic Monthly, 228*(3), 43–64.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. Simon & Schuster.

Hilgard, E. R. (1989). The early years of intelligence measurement. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 7–28). University of Illinois Press.

Holahan, C. K., Sears, R. R., & Cronbach, L. J. (1995). *The gifted group in later maturity*. Stanford University Press.

Holmgren, D. (2009). Lindquist, Everet Franklin. In D. Hudson, M. Bergman, & L. Horton (Eds.), *The biographical dictionary of Iowa*. University of Iowa Press Digital Editions.

Holzinger, K. J., & Clayton, B. (1925). Further experiments in the application of Spearman's prophecy formula. *Journal of Educational Psychology, 16*(5), 289–299. https://doi.org/10.1037/h0075199

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6*, 153–160. https://doi.org/10.1007/BF02289270

Huang, W. (2015, December 15). *Benjamin Wright, renowned psychometrician, 1926–2015*. UChicago News. https://news.uchicago.edu/story/benjamin-wright-renowned-psychometrician-1926-2015

Irribarra, D. T. (2021). *A pragmatic perspective of measurement*. Springer. https://doi.org/10.1007/978-3-030-74025-2

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39*(1), 1–123.

Jensen, A. R. (1980, April). Letter to Janice Scheuneman. Collection of the authors.

Jensen, A. R. (1998). *The G factor: The science of mental ability*. Praeger.

Jones, L. V. (2007). Some lasting consequences of the U.S. psychology programs in World Wars I and II. *Multivariate Behavioral Research, 42*(3), 593–608. https://doi.org/10.1080/00273170701382542

Jones, L. V., & Olkin, I. (2004). *The nation's report card: Evolution and perspectives*. Phi Delta Kappa Educational Foundation.

Kane, M. T. (2002). Inferences about variance components and reliability—generalizability coefficients in the absence of random sampling. *Journal of Educational Measurement, 399*(2), 165–181. https://doi.org/10.1111/j.1745-3984.2002.tb01141.x

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education; Praeger.

Kelley, T. L. (1921). The reliability of test scores. *Journal of Educational Research, 3*(5), 370–379. https://doi.org/10.1080/00220671.1921.10879169

Kelley, T. L. (1923). *Statistical method*. Macmillan.

Kelley, T. L. (1925). The applicability of the Spearman-Brown formula for the measurement of reliability. *Journal of Educational Psychology, 16*, 300–303. https://doi.org/10.1037/h0073506

Kelley, T. L. (1927). *Interpretation of educational measurements*. World Book.

Kelly, F. J. (1915). *The Kansas Silent Reading Test*. Kansas State Printing Plant.

Kendall, M. G. (1968). Studies in the history of probability and statistics. XIX Francis Ysidro Edgeworth, 1845–1926. *Biometrika, 55*(2), 269–275.

Kolen, M. (2021). History of test equating methods and practices through 1985. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 318–342). Routledge. https://doi.org/10.4324/9780367815318-14

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of reliability. *Psychometrika, 2*, 151–160. https://doi.org/10.1007/BF02288391

Larmer, B. (2014, December 31). Inside a Chinese test-prep factory. *New York Times Magazine*. https://www.nytimes.com/2015/01/04/magazine/inside-a-chinese-test-prep-factory.html

Lawrence, I., Rigol, G., Van Essen, T., & Jackson, C. (2004). A historical perspective on the content of the SAT. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 57–74). Routledge Falmer. https://doi.org/10.4324/9780203463932

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362–412). Princeton University Press.

Lemann, N. (2004). A history of admissions testing. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 5–14). Routledge Falmer. https://doi.org/10.4324/9780203463932

Levy, R., & Mislevy, R. J. (2021). A history of Bayesian inference in educational measurement. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement*: *Key advancements in theory, policy, and practice* (pp. 292–317). Routledge. https://doi.org/10.4324/9780367815318-13

Lindquist, E. F. (Ed.). (1951). *Educational measurement*. American Council on Education.

Lindquist, E. F. (1976, October 5). *Interview by James Beilman with Dr. Everet F. Lindquist* (Interview transcript No. XXII). University of Iowa Oral History Project, 1976–1977.

Linn, R. L. (Ed.). (1989). *Educational measurement* (3rd ed.). American Council on Education; Macmillan.

Lippmann, W. (1922a). The mental age of Americans. *The New Republic, 32*(412), 213–215.

Lippmann, W. (1922b). The mental age of Americans (Part II): Mystery of the "A" men. *The New Republic, 32*(413), 246–248.

Lippmann, W. (1922c). The mental age of Americans (Part III): Reliability of the intelligence tests. *The New Republic, 32*(414), 275–277.

Lippmann, W. (1922d). The mental age of Americans (Part IV): Abuse of the tests. *The New Republic, 32*(415), 297–298.

Lippmann, W. (1922e). The mental age of Americans (Part V): Tests of hereditary intelligence. *The New Republic, 32*(416), 328–330.

Lippmann, W. (1922f). The mental age of Americans (Part VI): A future for the tests. *The New Republic, 33*(417), 9–11.

Lippmann, W. (1923). The great confusion. *The New Republic, 33*(422), 145–146.

Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph No. 7). Psychometric Corporation.

Lord, F. M. (1954). Scaling. *Review of Educational Research, 24*, 375–393. https://doi.org/10.2307/1169042

Lord, F. M. (1968). *Some theory for tailored testing* (ETS Research Bulletin No. RB-68-38). ETS. https://doi.org/10.1002/j.2333-8504.1968.tb00562.x

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtsman (Ed.), *Computer-assisted instruction, testing, and guidance* (pp. 139–183). Harper & Row.

Lord, F. M. (1971a). Robbins–Monro procedures for tailored testing. *Educational and Psychological Measurement, 31*(1), 3–31. https://doi.org/10.1177/001316447103100101

Lord, F. M. (1971b). The self-scoring flexilevel test. *Journal of Educational Measurement, 8*(3), 147–151. https://doi.org/10.1111/j.1745-3984.1971.tb00918.x

Lord, F. M. (1971c). Tailored testing, an application of stochastic approximation. *Journal of the American Statistical Association, 66*(336), 707–711. https://doi.org/10.1080/01621459.1971.10482333

Lord, F. M. (1971d). A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement, 31*(4), 805–813. https://doi.org/10.1177/001316447103100401

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison–Wesley.

Lovie, P., & Lovie, A. D. (1996). Charles Edward Spearman, F.R.S. (1863–1945). *Notes and Records of the Royal Society of London, 50*(1), 75–88. https://doi.org/10.1098/rsnr.1996.0007

Lovie, S., & Lovie, P. (2010). Commentary: Charles Spearman and correlation: A commentary on "the proof and measurement of association between two things." *International Journal of Epidemiology, 39*(5), 1151–1153. https://doi.org/10.1093/ije/dyq183

Luecht, R. M., & Hambleton, R. K. (2021). Item response theory: A historical perspective and brief introduction to applications. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 232–262). Routledge. https://doi.org/10.4324/9780367815318-11

Mari, L., Wilson, M., & Maul, A. (2021). *Measurement across the sciences: Developing a shared concept system for measurement*. Springer. https://doi.org/10.1007/978-3-030-65558-7

Martha L. Stocking Swanson. (2006, December 29). *The Times of Trenton*. https://obits.nj.com/us/obituaries/trenton/name/martha-swanson-obituary?id=14042800

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174. https://doi.org/10.1007/BF02296272

McGrane, J. A., & Maul, A. (2020). The human sciences, models and metrological mythology. *Measurement, 152*, Article 107346. https://doi.org/10.1016/j.measurement.2019.107346

Mendel, J. G. (1866). Versuche über plflanzen-hybriden [Experiments on plant hybrids]. *Verhandlungen des Naturforschenden Vereines in Brünn, Bd. IV*, 3–47. https://doi.org/10.5962/bhl.title.61004

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*(10), 955–966. https://doi.org/10.1037/0003-066X.30.10.955

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education; Macmillan.

Messick, S., Beaton, A., & Lord, F. (1983). *National Assessment of Educational Progress reconsidered: A design for a new era* (Report No. NAEP-83-01). ETS.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge University Press. https://doi.org/10.1017/CBO9780511490040

Minton, H. L. (1990). Lewis M. Terman and mental testing: In search of the democratic ideal. In M. M. Sokal (Ed.), *Psychological testing and American society, 1890–1930* (pp. 95–112). Rutgers University Press.

Morant, G. M., & Welch, B. L. (1939). *A bibliography of the statistical and other writings of Karl Pearson*. Cambridge University Press.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176. https://doi.org/10.1177/014662169201600206

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Department of Education.

Otis, A. S. (1918a). An absolute point scale for the group measurement of intelligence. Part 1. *Journal of Educational Psychology, 9*(5), 239–261. https://doi.org/10.1037/h0072885

Otis, A. S. (1918b). An absolute point scale for the group measurement of intelligence. Part 2. *Journal of Educational Psychology, 9*(6), 333–348. https://doi.org/10.1037/h0074753

Otis, A. S., & Knollin, H. E. (1921). The reliability of the Binet scale and of pedagogical scales. *Journal of Educational Research, 4*(2), 121–142. https://doi.org/10.1080/00220671.1921.10879187

Pearson, K. (1896). Mathematical contributions to the theory of evolution: III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London, Series A, 187*, 253–318. https://doi.org/10.1098/rsta.1896.0007

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Series 5, 50*(302), 157–175. https://doi.org/10.1080/14786440009463897

Pearson, K. (1904). On the laws of inheritance in man: II. On the inheritance of mental and moral characters in man, and its comparison with the inheritance of physical characters. *Biometrika, 3*(2–3), 131–190. https://doi.org/10.1093/biomet/3.2-3.131

Pearson, K. (1907). *Mathematical contributions to the theory of evolution: XVI. On further methods of determining correlation*. Dulau & Co.

Pearson, K. (1934). Reply from Professor Karl Pearson. In *Speeches delivered at a dinner held in University College, London in honour of Professor Karl Pearson, 23 April 1934* (pp. 19–24). Cambridge University Press.

Pearson, K., & Moul, M. (1927). The mathematics of intelligence: I. The sampling errors in the theory of a generalized factor. *Biometrika, 19*(3–4), 246–291. https://doi.org/10.1093/biomet/19.3-4.246

Perez, C. (2002). Different tests, same flaws: Examining the SAT I, SAT II, and ACT. *The Journal of College Admission, 177*, 20–25.

Phelps, R. P. (2007). *Standardized testing primer*. Peter Lang.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika, 30*(1), 39–56. https://doi.org/10.1007/BF02289746

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability: Vol. 4. Contributions to biology and problems of medicine* (pp. 321–333. University of California Press.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19*(1), 49–57. https://doi.org/10.1111/j.2044-8317.1966.tb00354.x

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. MESA Press; University of Chicago. (Original work published 1960)

Reese, W. J. (2013). *Testing wars in the public schools: A forgotten history*. Harvard University Press.

Remmers, H. H., Shock, N. W., & Kelly, E. L. (1927). An empirical study of the validity of the Spearman–Brown formula as applied to the Purdue rating scale. *Journal of Educational Psychology, 18*(3), 187–195. https://doi.org/10.1037/h0072665

Rice, J. M. (1913). *Scientific management in education*. Publishers Printing Company.

Rice, J. M. (1897a). The futility of the spelling grind: I. *Forum, 23*, 163–172.

Rice, J. M. (1897b). The futility of the spelling grind: II. *Forum, 23*, 409–419.

Robinson, D. (2005). Profiles in research: Howard Wainer. *Journal of Educational and Behavioral Statistics, 30*(4), 465–476. https://doi.org/10.3102/10769986030004465

Robinson, D. H., & Wainer, H. (2006). Profiles in research: Arthur Jensen. *Journal of Educational and Behavioral Statistics, 31*(3), 327–352. https://doi.org/10.3102/10769986031003327

Ruch, G. M., Ackerson, L., & Jackson, J. D. (1926). An empirical study of the Spearman–Brown formula as applied to educational test material. *Journal of Educational Psychology, 17*(5), 309–313. https://doi.org/10.1037/h0075480

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Psychometric Society.

Samejima, F. (1972). *A general model for free-response data* (Psychometric Monograph No. 18). Psychometric Society.

Samelson, F. (1990). Was early mental testing: (a) racist inspired, (b) objective science, (c) a technology for democracy, (d) the origin of the multiple-choice exams, (e) none of the above? (Mark the RIGHT answer). In M. M. Sokal (Ed.), *Psychological testing and American society, 1890–1930* (pp. 113–127). Rutgers University Press.

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1999). *CATBOOK computerized adaptive testing: From inquiry to operation* (Study Note No. 99-02). U.S. Army Research Institute for the Behavioral and Social Sciences.

Shields, J., Hanser, L. M., & Campbell, J. P. (2001). A paradigm shift. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 19–28). Taylor & Francis.

Sitgreaves, R. (1963). Book review: *Probabilistic models for some intelligence and attainment tests* by G. Rasch. *Psychometrika, 28*(2), 219–220. https://doi.org/10.1007/BF02289619

Sokal, M. M. (1990). James McKeen Cattell and mental anthropometry: Nineteenth-century science and reform and the origins of psychological testing. In M. M. Sokal (Ed.), *Psychological testing and American society, 1890–1930* (pp. 21–45). Rutgers University Press.

Spearman, C. (1904a). "General intelligence" objectively determined and measured. *The American Journal of Psychology, 15*(2), 201–293. https://doi.org/10.2307/1412107

Spearman, C. (1904b). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101. https://doi.org/10.2307/1412159

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *The American Journal of Psychology, 18*, 161–169.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Spearman, C. (1930). Autobiography. In C. Murchison (Ed.), *A history of psychology in autobiography* (Vol. 1, pp. 299–333). Clark University Press.

Staff, Psychology Branch, Office Air Surgeon. (1944). The aviation cadet qualifying examination of the Army Air Forces. *Psychological Bulletin, 41*(6), 385–394. https://doi.org/10.1037/h0059734

Stanley, J. C. (1966). Rice as a pioneer educational researcher. *Journal of Educational Measurement, 3*(2), 135–139. https://doi.org/10.1111/j.1745-3984.1966.tb00869.x

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). Wiley.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Belknap Press.

Stigler, S. M. (1999). *Statistics on the table: The history of statistical concepts and methods*. Harvard University Press.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*(1), 57–75. https://doi.org/10.2307/116534

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of item in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163–182). Springer. https://doi.org/10.1007/0-306-47531-6_9

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201–210. https://doi.org/10.1177/014662168300700208

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*(3), 277–292. https://doi.org/10.1177/014662169301700308

Stocking, M. L., & Swanson, D. (1998). Optimal design of item banks for computer adaptive testing. *Applied Psychological Measurement, 22*(3), 271–279. https://doi.org/10.1177/01466216980223007

Stocking, M. L., Swanson, L., & Pearlman, M. (1991a). *Automatic item selection (AIS) methods in the ETS testing environment* (ETS Research Memorandum No. RM-91-05). ETS.

Stocking, M. L., Swanson, L., & Pearlman, M. (1991b). *Automated item selection using item response theory* (ETS Research Report No. RR-91-09). ETS. https://doi.org/10.1002/j.2333-8504.1991.tb01375.x

Stonehill, R. M., & Anderson, J. I. (1982). *An evaluation of ESEA Title I—program operations and educational effects. A report to Congress*. U.S. Department of Education, Budget, and Evaluation, Office of Planning.

Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarsfeld, P. L., Star, S. A., & Clausen, J. A. (1950). *Measurement and prediction*. Princeton University Press.

Terman, L. M. (1917). The intelligence quotient of Francis Galton in childhood. *The American Journal of Psychology, 28*(2), 209–215. https://doi.org/10.2307/1413721

Terman, L. M. (1922). The great conspiracy, or the impulse imperious of intelligence testers, psycho-analyzed and exposed by Mr. Lippmann. *The New Republic, 33*(421), 116–120.

Terman, L. M., Baldwin, B. T., Bronson, E., DeVoss, J. C., Fuller, F., Goodenough, F. L., Kelley, T. L., Lima, M., Marshall, H., Moore, A. H., Raubenheimer, A. S., Ruch, G. M., Willoughby, R. L., Wyman, J. B., & Yates, D. H. (1925). *Mental and physical traits of a thousand gifted children: Vol. 1. Genetic studies of genius.* Stanford University Press.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501–519. hhttps://doi.org/10.1007/BF02302588

Thissen, D., & Steinberg, L. (2020). An intellectual history of parametric item response theory models in the twentieth century. *Chinese/English Journal of Educational Measurement and Evaluation, 1*(1), 23–39. https://doi.org/10.59863/GPML7603

Thorndike, E. L. (1903). *Educational psychology*. Lemcke & Buechner. https://doi.org/10.59863/GPML7603

Thorndike, E. L. (1913). *Educational psychology, Vol. 1: The original nature of man.* Teacher's College, Columbia University.

Thorndike, E. L. (1919). *An introduction to the theory of mental and social measurements*. Columbia University, Teachers College.

Thorndike, R. L. (1947). *Research problems and techniques* (Vol. 3). Army Air Forces, U.S. Government Printing Office.

Thorndike, R. L. (1949). *Personnel selection; Test and measurement techniques*. John Wiley.

Thorndike, R. L. (Ed.). (1971). *Educational measurement* (2nd ed.). American Council on Education.

Thorndike, E. L., Lay, W., & Dean, P. R. (1909). The relation of accuracy in sensory discrimination to general intelligence. *American Journal of Psychology, 20*, 364–369.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16*(7), 433–451. https://doi.org/10.1037/h0073357

Thurstone, L. L. (1927). The unit of measurement in educational scales. *Journal of Educational Psychology, 18*(8), 505–524. https://doi.org/10.1037/h0075524

Thurstone, L. L. (1931). *The reliability and validity of tests*. Edwards Brothers.

Tinker, M. A. (1932). Wundt's doctorate students and their theses 1875–1920. *The American Journal of Psychology, 44*(4), 630–637. https://doi.org/10.2307/1414529

Torgerson, W. S. (1958). *Theory and methods of scaling*. Wiley.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11*, 1–13. https://doi.org/10.1007/BF02288894

Urban, W. J. (2010). *More than science and* Sputnik*: The National Defense Education Act of 1958*. University of Alabama Press.

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27–52). Kluwer. https://doi.org/10.1007/0-306-47531-6_2

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. Springer. https://doi.org/10.1007/978-1-4757-2691-6

Wainer, H. (2000). Introduction and history. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 1–22). Lawrence Erlbaum Associates.

Wainer, H. (2005). *Graphic discovery: A trout in the milk and other visual adventures*. Princeton University Press.

Wainer, H., & Kiely, G. L. (1987). Item clusters in computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185–201. https://doi.org/10.1111/j.1745-3984.1987.tb00274.x

Wainer, H., & Robinson, D. H. (2006). Profiles in research: R. Darrell Bock. *Journal of Educational and Behavioral Statistics, 31*(1), 101–122. https://doi.org/10.3102/10769986031001101

Wainer, H., & Robinson, D. H. (2009). Profiles in research: Linda S. Gottfredson. *Journal of Educational and Behavioral Statistics, 34*(3), 395–427. https://doi.org/10.3102/1076998609339366

Wallace, A. R. (1858). On the tendency of varieties to depart indefinitely from the original type. *Zoological Journal of the Linnean Society, 3*, 46–50.

Webster, L. (1998). David Andrich: A genius from down under. *Popular Measurement, 1*, 26. https://www.rasch.org/pm/pm1-26.pdf

Weinland, T. P. (1973). A history of the I.Q. in America, 1890–1941. *Dissertation Abstracts International Section A: Humanities and Social Sciences, 33*(10-A), 5666.

Weiss, D. J. (Ed.). (1978). *Proceedings of the 1977 computerized adaptive testing conference*. University of Minnesota.

Weiss, D. J. (Ed.). (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. Academic Press.

Wissler, C. (1901). The correlation of mental and physical tests. *Psychological Review: Monograph Supplements, 3*(6), i–62. https://doi.org/10.1037/h0092995

Wolf, T. A. (1973). *Alfred Binet*. University of Chicago Press.

Wood, B. D. (1926). Studies in achievement tests. Part III. Spearman–Brown reliability predictions. *Journal of Educational Psychology, 17*(4), 263–269. https://doi.org/10.1037/h0063759

Wood, D. A. (1962). *Louis Leon Thurstone: Creative thinker, dedicated teacher, eminent psychologist*. ETS.

Wright, B. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 ETS Invitational Conference on Testing Problems* (pp. 85–101). ETS.

Wright, B. (1992, April 20–24). *IRT in the 1990s: Which models work best? 3PL or Rasch?* [Paper presentation]. American Educational Research Association Annual Meeting. San Francisco, CA, United States. https://www.rasch.org/rmt/rmt61a.htm

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). American Council on Education; Praeger.

Yerkes, R. M. (Ed.). (1921). *Psychological examining in the army* (Vol. XV). National Academy of Sciences.

Yoakum, C. S., & Yerkes, R. M. (Eds.). (1920). *Mental tests in the American army*. Sidgwick & Jackson. https://doi.org/10.1037/11054-000

Yule, G. U. (1897). On the significance of Bravais' formulæ for regression, &c., in the case of skew correlation. *Proceedings of the Royal Society of London, 60*, 359–367. https://doi.org/10.1098/rspl.1896.0075

Zenderland, L. (1990). The debate over diagnosis: Henry Herbert Goddard and the medical acceptance of intelligence testing. In M. M. Sokal (Ed.), *Psychological testing and American society, 1890–1930* (pp. 46–74). Rutgers University Press.

Zenderland, L. (1998). *Measuring minds: Henry Herbert Goddard and the origins of American intelligence testing*. Cambridge University Press.

Zhang, H., & Luo, F. (2020). The development of psychological and educational measurement in China. *Chinese/English Journal of Educational Measurement and Evaluation, 1*, 56–64. https://doi.org/10.59863/BUAI8988

NOTES

1. The first and second wranglers that same year are decidedly less well known: Andrew James Campbell Allen and George Francis Walker. Other well-known scholars who performed well in the examinations but failed to achieve first place include Alfred North Whitehead, John Venn, Bertrand Russell, Thomas Malthus, John Maynard Keynes, and R. A. Fisher.
2. Several other formulations of this equation may be more familiar to the reader, including $r_{xy} = \frac{\Sigma(xy)}{\sqrt{\Sigma x^2 \Sigma y^2}}$.
3. This statement obviously reflects Pearson's enthusiasm for eugenics, but it should be noted that it was made years before the atrocities of the Holocaust were widely known.
4. The form of the equation provided in the text is the one presented by Spearman. The reader may be more familiar with the following equivalent formulation in which x and y replace p and q, $r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}$.
5. The reader may be familiar with the following formulation in which r_{11} is the reliability of the original test and r_{KK} is the reliability of a test lengthened by a factor of K, $r_{KK} = \frac{K r_{11}}{1 + (K-1) r_{11}}$.
6. Clauser (2021) provided a more detailed presentation of the material in this section.
7. These papers were originally published in 1905, 1908, and 1911. In 1916, at Goddard's instigation, they were translated and published in a single volume. In what follows, the chronology of development will be based on the original publication dates, but specific references will be to the translation.
8. Some previous authors explicitly state that Binet and Simon may have been the first psychologists to work with item characteristic curves (e.g., Hambleton & Swaminathan, 1985). This requires a broad definition of *item characteristic curve*. Binet and Simon examined the probability of a correct response as a function of age rather than as a function of estimated ability. In this context, it is also worth noting that it does not appear that Binet and Simon actually plotted their results; instead, they presented the results in tabular form. That said, several of the tables in

their paper on application of the new methods present scores as a function of age and could scarcely be more informative if they were converted to graphic form.

9. Our copy was signed by Goddard in 1949 with a note supporting such beneficial outcomes claiming that "Deborah Kallikak still lives, not in story but in fact. She is now 58 years old in good health and enjoying life . . . across the street from the Vineland Training School where she was living when this book was written."
10. The idea of forming a quotient by dividing mental age by chronological age was originally recommended by the German psychologist William Stern, but the widespread use and its transfer to a three-digit scale—with 100 representing a 1:1 ratio of mental age to chronological age—is the result of Terman's work.
11. Not to be mistaken for the publishers of the *World Book Encyclopedia*.
12. The Terman study may be the longest running psychological study. The Fels study ran longer—from 1929 to 2018—however, this study was primarily focused on physiological measurements.
13. In addition to Terman's effort to follow individuals into adulthood to examine how their exceptional IQs impacted their lives, he apparently thought it was possible to estimate the IQs of historical figures; he published a paper in which he estimated Francis Galton's childhood IQ by evaluating his childhood mental age based on records of his childhood behavior. He concluded that his IQ must have been around 200 (Terman, 1917).
14. Brigham subsequently changed his mind about the usefulness of intelligence tests for making racial comparisons. In Brigham (1930), he made this change of opinion clear, stating in reference to *A Study of American Intelligence*, "This review has summarized some of the more recent test findings which show that comparative studies of various national and racial groups may not be made with existing tests, and which show, in particular, that one of the most pretentious of these comparative racial studies—the writer's own—was without foundation" (p. 165).
15. Even Helen Keller supported the eugenics movement. In 1915, she wrote a letter to *New Republic* supporting a doctor who withheld treatment from a severely disabled baby. In that letter she advocated the formation of juries of physicians to make such decisions, stating, "A mental defective . . . is almost sure to be a potential criminal. The evidence before a jury of physicians considering the case of an idiot would be exact and scientific. . . . They would act only in cases of true idiocy, where there could be no hope of mental development." She went on to say, "Conservatives ask too much perfection of these new methods and institutions, although they know how far the old ones have fallen short of what they were expected to accomplish. We can only wait and hope for better results as the average of human intelligence, trustworthiness and justice arises."
16. Readers may be more familiar with this formulation in which the subscript *t* signifies true score and the subscript *x* signifies total score: $r_{12} = \frac{\sigma_t^2}{\sigma_x^2}$.

17. It is worth noting that a paper by Cyril Hoyt (1941) published in *Psychometrika* a decade earlier (and cited by Cronbach) also demonstrated that the KR20 represented the mean of all possible split-half estimates. Hoyt's paper also provided a link to generalizability theory by producing the estimate based on analysis of variance.
18. The suggestion that error of measurement might be a more useful metric than a reliability coefficient did not originate with Cronbach. Otis and Knollin (1921) and Kelley (1921) both made this argument. As we noted, the histories of the classical test theory and generalizability theory perspectives are intertwined.
19. This is a simplified description of generalizability theory. For a more complete description, see Cronbach et al. (1972) or Brennan (2001a). Brennan (2021) specifically presented aspects of the history of generalizability theory.
20. In addition to this, Edgeworth also published two papers on correlation years before Pearson's paper and he coined the term *coefficient of correlation*.
21. In the second edition of *Educational Measurement*, William Angoff wrote a chapter titled "Scales, Norms, and Equivalent Scores" (Angoff, 1971). Angoff viewed scaling as the process of defining the scale on which scores are to be reported.
22. This relationship would have been familiar to Fechner through his research on electricity. Georg Ohm had written that the loss of current would be a logarithmic function of the length of the wire: $V = m \text{ Log } (1 + x)$.
23. Lord also proposed the less well-known four-parameter IRT model.
24. In this context, it is worth noting that Lord was substantially focused on models that were applicable to multiple-choice questions. Rasch's interests were more general. Rasch (1960/1980) presented a model for reading accuracy, a model for reading speed, and a model for intelligence tests. In reviewing Rasch's book for *Psychometrika*, Sitgreaves (1963) noted that the first two of these models are more fully developed than the third.
25. Wright (1968) was based on a paper presented at ETS in 1967, but Wright did not actually discuss the model in that paper—it is included as a footnote, which may have been added later.
26. Thissen and Steinberg (2020) attributed the use of the term one-parameter model to Wainer.
27. The research reported in this volume began with his study of spelling in 1895. Work on mathematics and language followed in 1901 and 1903, respectively.
28. Kelley (1927) considered this an inappropriate practice. He commented, "On the average, in the neighborhood of .90 of the capacity measured by an all-round achievement battery score,—reading arithmetic, science, history, etc.—and the capacity measured by a general intelligence test is one and the same. If a comprehensive educational achievement test and a general intelligence test each give 'fairly reliable' total scores, each would need to be more than ten times as

long to yield equally reliable measures of the difference between the educational achievement and intelligence scores" (pp. 21–22). It may be that the intentions behind these approaches to testing differ more than the results.

29. SAT was originally an initialism representing Scholastic Aptitude Test. The name was subsequently changed to Scholastic Assessment Test and is now officially SAT.
30. Similar to the SAT, ACT originally was an initialism for American College Testing. ACT is now the official name of the test.
31. For more information on the evolution of NAEP, see Jones and Olkin (2004).
32. This information is inferred from the names of the authors. In the case of K. W. Vaughn, this inference was not possible and we were unable to find other specific information about this author.