Classical Test Theory in Historical Perspective

Ross E. Traub
The Ontario Institute for Studies in Education of the University of Toronto

What were the historic origins of classical test theory? What have been major milestones in its development?

Classical test theory is an emanation of the early 20th century. It was born of a ferment the ingredients of which included three remarkable achievements of the previous 150 years: A recognition of the presence of errors in measurements, a conception of that error as a random variable, and a conception of correlation and how to index it. Then in 1904 Charles Spearman showed us how to correct a correlation coefficient for attenuation due to measurement error and how to obtain the index of reliability needed in making the correction. Spearman's demonstration marked the beginning, as I see it, of classical test theory. Subsequently, the framework of classical theory was elaborated and refined by Spearman, George Udny Yule, Truman Lee Kelley, and others over the quarter century or so following 1904. Another milestone was laid in 1937 with the publication of the Kuder-Richardson formulas. This event was followed, shortly thereafter, by the idea of lower bounds to reliability and the framework for enhanced understanding found in the work of Louis Guttman. The culmination of classical test theory was realized in the systematic treatment it received from Melvin Novick (1966; Lord & Novick, 1968).

What follows is an attempt to add a little flesh to the bare bones of the foregoing outline. Before doing so, however, I should be clear about what I take classical test theory to be.

Classical Theory Circumscribed

Classical[1] test theory is founded on the proposition that measurement error, a random latent variable, is a component of the observed score random variable. The latter variable is realized in the measurements that may be taken of a characteristic of the persons[2] in some more-or-less well-defined population. Add to this proposition (a) a result that is true by construction—namely, that the error variable has zero covariance with that other latent component of observed measurements, the true score variable (Lord & Novick, 1968)—and (b) a crucial assumption that the error component of a measure is independent of the error components of other measures, either of the same characteristic or of different characteristics, and one is able to prove the basic theorems of classical test theory. Essential adjuncts to the theory are the ancillary assumptions invoked and experimental procedures followed in estimating coefficients of reliability and standard errors of measurement, two of classical test theory's basic results.

The Zeitgeist

A prominent feature of the zeitgeist at the turn of the 20th century was the notion of errors in scientific observations. As early as the 17th century, apparently, "Galileo had . . . reasoned that errors of observation are distributed symmetrically and tend to cluster around their true value" (Read, 1985, p. 348). "By the eighteenth century," according to Churchill Eisenhart (1983a),

> the practice of taking the arithmetic mean [of a set of observations] "for truth" had become fairly widespread. Nonetheless, Thomas Simpson (1710–1761) wrote to the president of the Royal Society in March 1755 that some persons of considerable note maintained that one single observation, taken with due care, was as much to be relied on as the mean of a great number. As this appeared to him to be a matter of much importance, he said that he had a strong inclination to ascertain whether by the application of mathematical principles the utility and advantage of the practice might be demonstrated. Measurements and functions of measurements, such as their arithmetic means, are not amenable to mathematical theory, however, as long as individual measurements are regarded as unique entities, that is as *fixed numbers* $y_1, y_2, \ldots$ A mathematical theory of measurements, and of functions of measurements, is possible only when particular measurements $y_1, y_2, \ldots$ are regarded as instances characteristic of hypothetical measurements $Y_1, Y_2, \ldots$ that might have been, or might be, yielded by the same measurement process under the same circumstances. Consequently, Simpson hypothesized that the respective chances of the different errors to

Ross E. Traub is a Professor at the Ontario Institute for Studies in Education, OISE, University of Toronto, 252 Bloor St. W., Toronto, Ontario, Canada M5S 1V6. His specializations are test theory and educational assessment.

> which any single observation is subject could be expressed as a discrete probability distribution of error. . . ." (p. 531)

By the beginning of the 19th century, astronomers had come to recognize errors in observations as an area worthy of research. Carl Friedrich Gauss derived the normal distribution while trying to prove that the mean of many observations of an unknown quantity, such as a parameter of the orbit of a planet, is the most likely value of that quantity (Eisenhart, 1983b; Read, 1985).[3] In other work—for example, that of Friedrich Wilhelm Bessel[4]—errors in astronomical observations were thought to be composed of many independent, elementary parts. The distribution of the sum of these errors was shown to be normal under various assumptions, hence the references, commonplace in the 19th century, to the law of errors in observations (Eisenhart, 1983b; Read, 1985). Later in the 19th century, however, as noted by O. B Sheynin (1968),

> [the normal distribution,] as far as the theory of errors is concerned, had been almost forgotten and even the concept of random errors of observation . . . [had become] divorced from the concept of random quantities in the theory of probability. . . . A qualifying remark should be added: . . . in the second half of the 19th century a definition of a random quantity as "dependent on chance" and possessing a certain law of distribution had become . . . natural. . . . As to random errors, these were usually taken to be errors with certain probabilistic properties, their specific distribution . . . being not so important. . . . It seems that Vassiliev (1885, *Theory of Probabilities*, in Russian, a lithographic edition, Kazan, p. 133) was the first who definitely held that random errors of observation are to be ranked among random quantities. (pp. 236–237)

We might expect the notion of correlation, that other essential feature of Spearman's zeitgeist, to have emerged from a study of the distribution of joint errors. Indeed, several individuals separately derived expressions resembling the "ordinates of the probability surface of normal correlation for two variables" (Walker, 1929, p. 93). These individuals included Robert Adrain, an Irishman who emigrated to the U. S. and taught at Rutgers, Columbia College, and the University of Pennsylvania and published such an equation in 1808; Pierre-Simon, Marquis de LaPlace published such an equation in 1810; Giovanni Antonio Amedeo Plana, professor of astronomy at Turin, published an equation in 1812; Gauss in 1823; and Auguste Bravais, professor of astronomy at Lyons and professor of physics at Paris, in 1846. But none of these scholars seems to have interpreted the cross-product term in the exponents of their expressions as an indicator of covariation or correlation. This enormous idea was left to Francis Galton. Although he did not actually derive the formula for the bivariate normal distribution—it can be argued that this was done by a Cambridge mathematician named J. D. Hamilton Dickson, to whom Galton gave the relevant concepts[5]—Galton did originate the basic ideas we associate with the bivariate distribution. The notion of the scatterplot, on which is displayed the two regression lines, can be found in Galton's 1885 presidential address to the Anthropological Section of the British Association, subsequently an article published in 1886 in the *Journal of the Anthropological Institute* under the title "Regression Towards Mediocrity in Hereditary Stature."[6] The term *correlation* was first used in a technical sense 2 years later in an article entitled "Co-Relations and Their Measurement" (Galton, 1888). Galton, however, used the symbol r to refer to *reversion* or *regression*. Credit for calling r the *coefficient of correlation* belongs to Francis Y. Edgeworth, and dates from 1892 (Boring, 1957).[7]

Karl Pearson contributed important mathematical work in support of his friend Galton's coefficient. For example, Pearson (1896) proved that the best value of r is the covariance divided by the product of the standard deviations.

Defining Achievements in the History of Classical Test Theory

We find ourselves at the very beginning of the 20th century, a time when the idea of errors in measurements had become widely accepted. Also, the coefficient of correlation was well established as an important statistical concept, although Pearson's product moment expression had not yet gained acceptance as the computational formula of choice. In addition, we encounter around this time the first scientific publications in which coefficients of correlation were the principal results of the research. For example, Karl Pearson, who, like Galton, had an abiding interest in eugenics, investigated the correlation between characteristics of pairs of brothers (Pearson, 1904; Pearson & Lee, 1903). Pearson found, to his astonishment, that the magnitude of this correlation was about 0.5, regardless of characteristic, physical or psychical. So this was the time, and this was the context for the first of the defining achievements in the history of classical test theory—the correction of a correlation coefficient for attenuation due to measurement error. In what follows, I describe this and four other milestones marking the development of classical test theory.

Spearman's correction for attenuation. Spearman was a psychologist. At the turn of the 20th century, he was just beginning his study of intelligence. In the course of this early work, he had discovered that independent measurements of a psychical characteristic of a person—for example, mental ability—vary in a random—Spearman called it an "accidental" (1904, p. 89)—fashion from one measurement trial to another. In other words, the coefficient of correlation between such independent measurements for a group of persons is not perfect. Spearman then had the insight that the absolute value of the coefficient of correlation between the measurements for any pair of variables must be smaller when the measurements for either or both variables are influenced by accidental variation than it would otherwise be. Eliminating this attenuating effect of accidental error was just one of the matters, albeit in retrospect the most important of those matters, addressed by Spearman in his 1904 article, entitled "The Proof and Measurement of

Association Between Two Things" and published in the *American Journal of Psychology*.

Spearman's article angered Karl Pearson. The reason was that Spearman had had the temerity to challenge Pearson's conclusion that the coefficient of correlation between pairs of brothers was 0.5 for psychic characteristics, just as it had been found to be for physical characteristics. Using reliability estimates from his own work, Spearman estimated the corrected correlation coefficient for mental ability to be 0.8.

Pearson was co-editor of a newly founded journal, *Biometrika*. He inserted the following petulant note in his 1904 article in that journal on the correlation between selected characteristics of sibling pairs.

> I hardly know whether it is needful to refer here to a recent article by Mr. C. Spearman . . . criticising my results for the similarity of inheritance in the physical and psychical characters. Without waiting to read my paper in full he seems to think I must have disregarded "home influences" and the personal equation of the school teachers. He proceeded to "correct" my results for the error of what he calls dilation on the double basis (i) of a formula invented by himself, but given without proof, and (ii) of his own experience that two observers' observations or measurements of the same series of two characters were such that the correlation between their determinations was .58 in one case and .22 in the other. The formula invented by Mr. Spearman for his so-called "dilation" is clearly wrong, for applied to perfectly definite cases, it gives values greater than unity for the correlation coefficient. As to his second basis, all I can say is that if the correlation between two observers of the same thing in Mr. Spearman's case can be as low as .22, he must have employed the most incompetent observers, or given them the most imperfect instructions, or chosen a character [more] suitable for random guessing than observation in the scientific sense. Mr. Spearman says that "it is difficult to avoid the conclusion that the remarkable coincidence announced between physical and mental heredity can be [nothing] more than mere accidental coincidence" (p. 98). I think I may safely leave him to calculate the odds for or against this most remarkable "mere accidental coincidence". . . . Perhaps the best thing at present would be for Mr. Spearman to write a paper giving algebraic proofs of all the formulae he has used; and if he did not discover their erroneous nature in the process, he would at least provide tangible material for definite criticism, which it is difficult to apply to mere unproven assertions. (Pearson, 1904, p. 160)

Stung into responding, Spearman published a proof of the correction for attenuation in 1907, again using the *American Journal of Psychology*. Subsequently, a proof in a form often encountered in present day textbooks on educational and psychological measurement was given by Spearman (1910) and also by William Brown (1910), both of whom ascribed it to Yule. This derivation stressed that the error components of all measures should be independent, and hence uncorrelated. Spearman's earlier "proof" had not emphasized this restriction (Walker, 1929).

Pearson was not the only critic of the correction for attenuation. In particular, Brown (1910) challenged it on the grounds that measurement error is not really random (accidental). Brown proposed a way of testing the equality of the covariances $S(x_1y_1)$ and $S(x_2y_2)$, where x_1, x_2, y_1, and y_2 are observed-score variables. Assuming x_1 is parallel to x_2 and y_1 is parallel to y_2, then both these covariances, according to classical theory, should equal $S(xy)$, where x and y are the true-score variables associated with $\{x_1, x_2\}$ and $\{y_1, y_2\}$, respectively. Brown (1913) reported results based on an application of this proposal, results he claimed did show that measurement errors are not accidental.

Spearman was aware of Brown's criticism, among others, and presciently responded as follows in his *British Journal of Psychology* article of 1910:

1. Spearman reiterated his position that of the two kinds of errors in measurement, "regular" and "accidental" (p. 273), the correction formula applies only to the latter. As regards Brown's contention that accidental errors can be correlated, Spearman observed that, if errors were indeed found to be linked (as might be the case, e.g., if a person were ill at the time of taking both tests x and y), then the investigator should employ a better experimental design.
2. It had been suggested according to Spearman (1910, p. 272) that investigators should make measurements so "efficient" that no correction would be needed. Spearman wondered how an investigator would know his measurements were efficient enough except by using the correction formula?
3. To Pearson's criticism that the correction could produce coefficients greater than one, Spearman countered that this might occur due to sampling error. He recommended (p. 277) the coefficient be set to one whenever this happened.

The Spearman-Brown formula. Coefficients of reliability are needed in order to apply the correction for attenuation. In his 1904 article, Spearman had assumed the availability of two independent measurements of both the characteristics for which a corrected correlation coefficient is desired. The breakthrough, apparently achieved independently by Spearman and Brown, to a formula by which to calculate a reliability coefficient from the two halves of just one composite measure was published in adjacent articles in a 1910 issue of *The British Journal of Psychology*. Brown's proof of the formula is the more elegant and bears the stamp of Yule.

During the second and third decades of the 20th century, numerous experiments were conducted to test predictions of the Spearman-Brown formula. A thoughtful review of publications emanating from this preoccupation of early psychometricians can be found in the notes on test theory prepared by Louis Leon Thurstone (1932).

The index of reliability and other results. It is a worthwhile experience, though in the late 20th cen-

tury a humbling one, to read the articles on test theory that were published between 1910 and 1925 by Spearman, Brown, and Kelley, among others. These documents contain a great many of the basic results of classical test theory. Kelley's 1923 text, *Statistical Methods*, included a compilation of these results in a section on reliability theory. Kelley also stated in this text the definition of reliability that he championed throughout his long career: the coefficient of correlation between "comparable tests" (p. 203). Kelley laid down three conditions for test comparability:

> (1) sufficient fore-exercise should be provided to establish an attitude or set, thus lessening the likelihood of the second test being different from the first, due to a new level of familiarity with the mechanical features, etc.; (2) the elements of the first test should be as similar in difficulty and type to those in the second, pair by pair, as possible; but (3) should not be so identical in word or form as to commonly lead to a memory transfer of correlation between errors. (p. 203)

Kelley was critical of Brown (1910) for using the term *reliability coefficient* to refer to the correlation between scores on repeated administrations of the same test. Brown's definition fails to meet the third of Kelley's conditions. (Kelley's conclusion: "Accordingly the repetition of a test to secure a reliability coefficient is to be deprecated" Kelley, 1923, p. 203).

An important result, used by Spearman in his 1910 proof of the prophecy formula, was the expression for the correlation between two composite measures in terms of the variances and covariances of the components. From this result, Abelson (1911, p. 314) derived the formula for what came to be known several years later as the *index of reliability*.[8] It seems this name was coined quite by accident. Kelley independently derived the formula for the index in 1916 and then in using it wrote that "the extent to which the grade determined by means of this test of forty words would correlate with the true spelling ability of the individual is probably an even more significant index of reliability" (p. 74). Walker reported that "[w]hen Munroe published the formula in his *Introduction to the Theory of Educational Measurements* (1923) he ... [set the phrase in capital letters], ascribed it to Kelley, and established *Index of Reliability* as a definite term"(1929, p. 118).

The Kuder-Richardson formulas. Writing the coefficient of correlation between two composite measures in terms of the variances and covariances of the measures' components made it possible to study the effects of the characteristics of item scores on the characteristics of total test scores. By 1936, Marion Richardson, in an article in the first volume of *Psychometrika*, had demonstrated several propositions for tests composed of discrete, dichotomously scored items. Invoking the assumption that all test items have equal variances—Richardson noted that this assumption is close to true for a wide range of item difficulty values—he showed that "the rejection of items with low item-test correlations raises the reliability of a test, if the number of items is held constant" (p. 72). Richardson also showed that "for tests of homogeneous [item] difficulty and constant length, the true variance is proportional to the average item intercorrelation" (p.75).

Given his work on the relationship between item and total test scores, it is perhaps not surprising to find Richardson a co-author, with Frederic Kuder, of the blockbuster article of 1937, the one containing the famous Formulas 20 and 21. (In a footnote to the article, Kuder and Richardson reported they had independently arrived at the results contained in the article. They apparently discovered this fact quite by accident, at which time they decided to publish the results jointly.) Kuder and Richardson began their article with a critique of existing approaches to the estimation of reliability. Of the test–retest method, they said,

> [Using] the same form gives, in general, estimates that are too high because of material remembered on the second application of the test. This memory factor cannot be eliminated by increasing the length of time between the two applications, because of variable growth in the function tested within the population of individuals. These difficulties are so serious that the method is rarely used. (p. 151)

Of the split-half approach, Kuder and Richardson concluded that "although the authors have made no actual count, it seems safe to say that most technicians use the split-half method of estimating reliability" (p. 151). They then observed that the number of splits possible for an n-item test (n being an even number) is $\frac{n!}{2\left[\left(\frac{n}{2}\right)!\right]^2}$, a number so large for any test of reasonable length that the reliability estimates from all possible split-halves of the test are very likely to vary considerably. So the issue for Kuder and Richardson was to devise a procedure by which the information in the item scores of a test could be used to produce a single estimate of reliability.

In deriving Formulas 20 and 21, Kuder and Richardson began by writing the following expression—their Equation 3—for the correlation of a test composed of n dichotomously scored items with a hypothetical second test of n such items:

$$r_{tt} = \frac{\sigma_t^2 - \sum_{i=1}^{n} p_i q_i + \sum_{i=1}^{n} r_{ii} p_i q_i}{\sigma_t^2}, \quad (3)$$

where σ_t^2 is the test variance, $\sum_{i=1}^{n} p_i q_i$ is the sum of the item variances, and $\sum_{i=1}^{n} r_{ii} p_i q_i$ is the sum of the item true score variances. The problem Kuder and Richardson set for themselves was that of estimating r_{ii}, the item reliability coefficient. They did so under various assumptions. Those leading to Equation 20 were that the matrix of interitem correlation coefficients has Rank 1 and that all interitem correlation coefficients are equal. Setting r_{ii} equal to r_{ij}, the interitem correlation coefficient, Kuder and Richardson derived the result:

$$\bar{r}_{ii} = \frac{\sigma_t^2 - \sum p_i q_i}{\left(\sum \sqrt{p_i q_i}\right)^2 - \sum p_i q_i}$$

$$= \frac{\sigma_t^2 - n\overline{pq}}{(n-1)\,n\overline{pq}},$$

which, when substituted in Equation 3, and simplified, gives

$$r_{tt} = \frac{n}{n-1}\left[\frac{\sigma_t^2 - n\overline{pq}}{\sigma_t^2}\right].$$

Formula 21 was derived under the additional assumption of equal item difficulty indices.

Kelley criticized the KR formulas in a 1942 article in *Psychometrika* on the grounds that they are valid only for tests "with unity of purpose"—that is, for tests composed of items that share just one factor in common. Reiterating his long-standing advocacy of the parallel-test design for reliability estimation, Kelley went on to say "we conclude that a belief that two or more measures of a mental function exist is prerequisite to the concept of reliability, and, further, not only that they exist but that they are available before a measure of reliability is possible" (p. 76). As a further challenge to the KR formulation, Kelley demonstrated that a test with zero interitem covariances could produce a reasonable correlation with a "similar" (p. 81) test, even though its KR_{20} index would be zero.

It is obvious, a half-century later, that Kelley's view did not prevail. The KR formulations quickly received widespread acceptance, abetted in part by the publication of an article by Paul Dressel (1940). Dressel showed that, when all the items of a test intercorrelate perfectly and all item variances are equal, KR_{20} attains the value of 1; otherwise, it is less. He further demonstrated that KR_{20} can take values less than 0. Dressel also increased the applicability of KR_{20} by deriving a version for tests to which the correction for guessing is applied.

Lower bounds to reliability. Perhaps it was Dressel's demonstration that KR_{20} can be negative that marks the beginning of work on lower bounds to reliability. Alternatively, a case might be made for Philip Rulon's (1939) article in the *Harvard Educational Review*. Rulon introduced the notion, not the name, of essentially tau-equivalent test halves. The halves of a test are essentially tau equivalent if, for any examinee in the test population, the true score on one test half differs from the true score on the other half by a constant, which is the same for all examinees. Also, under the essentially tau-equivalent assumption, as distinct from the parallel test-halves assumption, the error variance for an examinee on one test half is not necessarily equal to the error variance for the examinee on the other half (see Lord & Novick, 1968, p. 50).

Whatever the wellspring of work on lower bounds, Louis Guttman (1945) published the first article, as far as I know, in which lower bounds to reliability are explicitly derived. But this *Psychometrika* article is important for a reason other than the lower bounds it contains. Guttman also offered a theoretical framework within which to treat, if not actually to reconcile, the antagonistic views of Brown and Kelley regarding how the reliability coefficient should be estimated. Guttman did this by first identifying three sources of variation in test responses—persons, items, and trials. Guttman defined error variance exclusively in terms of variation in responses over the universe of trials. This definition leads to a proof of total test variance as the sum of true-score and error variance, without the need to assume zero covariance of true and error scores. (The latter assumption lies at the heart of Yule's proof of the correction for attenuation.) Defining the reliability coefficient "as the complement of the ratio of error variance to total [test] variance" (p. 257), Guttman then went on to demonstrate (pp. 267–268) that the reliability coefficient can be estimated as the correlation between the test scores for a group of examinees on two "experimentally independent" (p. 264) trials of a test. Given the results from only one trial, however, Guttman showed that the best result possible is an estimate of a lower bound to reliability. This result was shown to rest on the assumption

> that the errors of observation are independent between items and between persons over the *universe of trials*. In the conventional [Yule] approach, independence is taken over *persons* rather than trials, and the problem of observability from a single trial is not explicitly analyzed. (p. 257)

Guttman derived six lower bounds to reliability, of which three are noted here. One of these is a generalization of KR_{20} to tests composed of items scored on any scale, dichotomous or otherwise. He labeled this index Γ_3. Subsequently, Γ_3 became better known as coefficient alpha (Cronbach, 1951). Two of the other lower bounds to reliability were labeled, not surprisingly, Γ_1 and Γ_2, with the former typically smaller than alpha, the latter typically larger. Research on lower bounds to reliability constitutes a small but still active line of psychometric research.

Formalization

Various attempts to formalize classical test theory have been made over the years. Already mentioned is the section on reliability in Kelley's (1923) *Statistical Method*. Another early work is that by Thurstone (1932). The next presentation of note is *Theory of Mental Tests* by Harold Gulliksen (1950). The culmination of such efforts as these was realized in the work of Melvin Novick (1966) and in the early chapters of *Statistical Theories of Mental Test Scores* (1968) by Frederic Lord and Novick.

Concluding Remarks

Several important topics from the realm of classical test theory have not been covered in this brief retrospective. Among them are the effects of range restriction on the magnitude of the reliability coefficient, the application of analysis of variance to the study of measurement error and reliability (this before the advent of generalizability theory), and the modeling of test data generally addressed under the topic of congeneric models. Clearly, there is more to classical test theory, and its history, than the work reviewed in this article.

Lest we leave this topic thinking classical test theory an unduly important area of research in the history of empirical research in psychology and education, we will find it salutary to reflect on the following

remarks from the preface of the 1932 edition of Thurstone's notes on test theory:

> Since this volume is devoted to the validity and reliability concepts in their applications to mental tests and related correlational procedures, it is only fair to say that personally I do not believe that these correlational methods and particularly the reliability formulae have been responsible for much that can be called fundamental, important or significant in psychology. On the contrary, the correlational methods have probably stifled scientific imagination as often as they have been of service. As tools in their proper place they are useful but as the central theme of mental measurement they are rather sterile.

Notes

This article is a revised version of a paper presented at the 1996 Annual Meeting of the National Council on Measurement in Education (Session B1), New York.

[1] Something *classical*, according to *Webster's New Collegiate Dictionary*, is accepted as standard and authoritative, as distinguished from novel or experimental: viz. classical physics.

[2] Classical test theory applies to the measurements of a characteristic of the members of any collection of objects whatsoever. In particular, it is not restricted, as the wording here might imply, to measurements of human characteristics.

[3] Although the normal distribution is commonly referred to as the Gaussian distribution, priority in its formulation rightfully belongs to Abraham de Moivre, who obtained it in 1733 (Read, 1985).

[4] In the 18th and early 19th centuries, astronomers were required to make difficult judgments, based on a combination of auditory and visual cues, in order to time stellar transits. A well-known story from the history of science (Boring, 1957) is the firing in 1796 of Kinnebrook, an assistant to Maskelyne, the Astronomer Royal of England. Kinnebrook was relieved of his job for giving inaccurate readings of stellar transits. Although he had provided readings in agreement with Maskelyne's 18 months prior to his dismissal, the hapless Kinnebrook by August 1795 had begun to give times that differed from Maskelyne's by one-half second. Subsequently, Kinnebrook's readings grew even more discrepant, so by the time of his firing they were almost a second later than Maskelyne's. This matter might not have attracted much interest had not Maskelyne recorded it in *Astronomical Observations at Greenwich* (1799). Seventeen years later, in a history of Greenwich Observatory published in German, Kinnebrook's tribulation came to the attention of Bessel, an astronomer at Königsberg. Bessel conducted a series of studies culminating in the notion of the personal equation—the name given the systematic difference in recording times found to characterize the stellar transits of almost any pair of astronomers. From the perspective of reliability theory, the personal equation itself was not a highly significant discovery, for it refers to systematic error, not the random error treated by reliability theory. What interests us, instead, is Bessel's finding that the personal equation itself is a variable quantity, one that differs from one pair of astronomers to another. This variation suggests random or accidental errors in observations, errors that, if neither controllable nor amenable to elimination, at the least demand an explanation grounded in a theory or a scientific law.

[5] Karl Pearson (1930) noted that Dickson did not actually write down the equation for the bivariate normal distribution but stopped one step short of doing so.

[6] According to Walker (1929, pp. 98–101), H. P. Bowditch published a two-way table of height and age for 24,000 Boston school boys in 1877 in a manuscript entitled *Growth of Children*. Although he described one of the regression lines of a bivariate distribution in his work, Bowditch neither produced both lines nor conceived of a measure of the relationship between the variables.

[7] Another of Galton's ideas was that the (normal) law of errors in observations might describe the frequency distributions of measurements of such human characteristics as mental ability. This idea was accepted readily enough by John Venn (1888), who wrote as follows: "That our mental qualities, if they could be submitted to accurate measurement, would be found to follow the usual Law of Error may be assumed without much hesitation" (p. 49). Venn expressed skepticism, however, over the idea that a normal distribution of mental measurements is yet another manifestation of the law of error:

> When we perform an operation ourselves with a clear consciousness of what we are aiming at, we may quite correctly speak of every deviation from this as being an error; but when Nature presents us with a group of objects of every kind, it is using a rather bold metaphor to speak in this case also of a law of error, as if she had been aiming at something all the time, and had like the rest of us missed her mark more or less in every instance. (p. 42)

[8] Walker (1929, p. 117) suggested the derivation of the formula may have been given to Abelson by Spearman. Abelson's article is, however, unclear on this point. The article contains two appendices. The first is described as having been "kindly supplied by Prof. Spearman" (p. 312). But the second appendix, which contains the index of reliability, bears no attribution to Spearman.

References

Abelson, A. R. (1911). The measurement of mental ability of "backward children." *British Journal of Psychology, 4*, 268–314.

Boring, E. G. (1957). *A history of experimental psychology* (2nd ed.). New York: Appleton-Century-Crofts.

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

Brown, W. (1913). The effects of "observational errors" and other factors upon correlation coefficients in psychology. *British Journal of Psychology, 6*, 223–235.

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

Dressel, P. L. (1940). Some remarks on the Kuder-Richardson reliability coefficient. *Psychometrika, 5*, 305–310.

Eisenhart, C. (1983a). Laws of error I: Development of the concept. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 4, pp. 530–547). Toronto: Wiley.

Eisenhart, C. (1983b). Laws of error II: The Gaussian distribution. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 4, pp. 547–562). Toronto: Wiley.

Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society, 45*, 135–145.

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

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255–282.

Kelley, T. L. (1916). A simplified method of using scaled data for purposes of testing. *School and Society, 4*, 34–37, 71–75.

Kelley, T. L. (1923). *Statistical method.* New York: Macmillan.
Kelley, T. L. (1942). The reliability coefficient. *Psychometrika, 7,* 75–83.
Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2,* 151–160.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.
Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3,* 1–18.
Pearson, K. (1896). Mathematical contributions to the theory of evolution—III. Regression, heredity and panmixia. *Philosophical Transactions, A, 187,* 252–318.
Pearson, K. (1904). On the laws of inheritance in man. II. On the inheritance of the mental and moral characters in man, and its comparison with the inheritance of physical characters. *Biometrika, 3,* 131–190.
Pearson, K. (1930). *The life, letters, and labours of Francis Galton. Vol. IIIA. Correlation, personal identification and eugenics.* Cambridge: The University Press.
Pearson, K, & Lee, A. (1903). On the laws of inheritance in man. I. Inheritance of physical characters. *Biometrika, 2,* 357–462.
Read, C. B. (1985). Normal distribution. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 6, pp. 347–359). Toronto: Wiley.
Richardson, M. W. (1936). Notes on the rationale of item analysis. *Psychometrika, 1*(1), 69–76.
Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review, 9,* 99–103.
Sheynin, O. B. (1968). On the early history of the law of large numbers. *Biometrika, 55,* 459–467.
Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101.
Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18,* 160–169.
Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295.
Thurstone, L. L. (1932). *The reliability and validity of tests.* Ann Arbor, MI: N. p.
Venn, J. (1888). *The logic of chance* (3rd ed.). London: Macmillan.
Walker, H. M. (1929). *Studies in the history of statistical method.* Baltimore: Williams & Wilkins.

A Perspective on the History of Generalizability Theory

Robert L. Brennan
University of Iowa

What psychometric and scientific perspectives influenced the development of G theory? What practical testing problems gave impetus to its adoption? What work remains to be done?

Overviews of various parts of the history of generalizability (G) theory are provided elsewhere. An indispensable starting point is the preface and parts of the first chapter of Cronbach, Gleser, Nanda, and Rajaratnam (1972) entitled *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* The Cronbach et al. monograph is still the most definitive treatment of G theory. Shavelson and Webb (1981) review the G theory literature from 1973–1980, and Shavelson, Webb, and Rowley (1989) cover additional contributions in the 1980s. A very brief historical overview is provided by Brennan (1983, 1992a, pp. 1–2). In addition, Cronbach (1976, 1989, 1991) offers numerous perspectives on G theory and its history. Cronbach (1991) is particularly rich with first-person reflections.

This historical overview is not intended to repeat everything already covered in published reviews, although a summary is provided. Parts of this article are based largely on my personal experience with G theory. Consequently, this article provides a somewhat idiosyncratic perspective on the history of G theory and what I perceive as unfinished work for the theory. Almost certainly, other reviewers would see the landscape somewhat differently.

Theory Development and Enabling Work

In discussing the genesis of G theory, Cronbach (1991) states:

> In 1957 I obtained funds from the National Institute of Mental Health to produce, with Gleser's

Robert L. Brennan is Lindquist Professor of Educational Measurement and Director of the Iowa Testing Programs, University of Iowa, 334A Lindquist Center, Iowa City, IA 52242. His specializations are generalizability theory, equating, and scaling.