2

Historical Precursors and Early Testing Theory

The present article…advocates a "Correlational Psychology," for the purpose of positively determining all psychicaltendencies in particular those which connect together the so-called mental tests with psychical activities of greater generality and interest.
—Spearman (1904b, p. 205)

Although Construct validity theory (CVT) did not formally enter the scene until psychology was into its eighth decade as an established independent discipline, it might be argued that the history of test-related validity as an area of scholarship is as old as the discipline itself. That is to say, the story of the history of disciplinary psychology is in many ways a story about psychological measurement, about attempts to represent psychological attributes quantitatively and then determine whether, and the extent to which, particular quantitative representations constitute "good" measurements of the psychological attribute in question. As was noted in the introductory chapter, early approaches to mental testing generally presupposed that quantitative representations of psychological attributes are, in principle, reasonable and legitimate and, thus, early testing theory was not really concerned with providing explicit theories

K. Slaney, *Validating Psychological Constructs*, Palgrave Studies in the Theory and History of Psychology, DOI 10.1057/978-1-137-38523-9_2

of psychological *measurement*. Rather, early testing theorists focused on two primary issues: (1) the impact of measurement error on indices of correlation; and (2) the extent to which concomitance among pairs of measurements revealed something fundamental to both measurements, namely, that they measured something in common. These two issues occupied much of the attention of early testing theorists and, in fact, continue to feature heavily in the technical psychometric literature.

As with many new theoretical-methodological frameworks, CVT was motivated in large part by the presence of ambiguities in the discourse, specifically, regarding how validity should be understood and, by extension, what ideally ought to be involved in validating claims based on psychological test data. Thus, in order to properly hinge the introduction of CVT to the psychological testing literature, it is necessary to document pre-CVT treatments of validity and validation as they pertain to quantitative psychological measurements and assessments. To this end, this chapter begins with a review of the work of British psychologist, Charles Spearman, who, in two separate papers published in 1904 (Spearman 1904a, b), articulated the foundations of what would later be called "classical test theory" and "factor theory," respectively. The contributions of other figures prominent in early psychometric theory will also be described. The chapter will wrap up with a description of classical conceptions of validity and the approaches to validation implied by it, and the differentiating of different aspects of validity that would anticipate a dramatic re-conceptualization of the concept toward the mid-twentieth century as older, "classical," conceptions would come under increasing scrutiny.

Test Theory for Mental Measurements

The publication of Darwin's *On the Origin of Species* in 1859 set in motion a new focus on individual variation with respect to particular traits and attributes in terms of which individuals within a species could be characterized and compared to one another. Darwin's half-cousin, Francis Galton embraced this idea and applied it in a rigorous way in his attempts to measure human traits and characteristics, both physical and mental. Borrowing from the psychophysical methods employed by Wundt and other early experimental psychologists, Galton developed the first battery of mental tests composed of "a peculiar assortment of sensory and motor measures," including reaction time and sensory discrimination tasks, which he administered to large numbers of individuals at his psychometric laboratory (Gregory 2004, p. 2). Borrowing from the methods of Quetelet, Galton rooted his explanations of the heritability of "eminence" in terms of correlations and deviations from averages with respect to measurements of a variety of mental attributes (Young 1923). Galton's legacy travelled to America with Cattell, who would become a vocal advocate of a psychology based on the measurement of individual differences in mental abilities (Cattell 1890). By the turn of the twentieth century, the use of mental tests for measuring a range of intellectual abilities and aptitudes was quickly taken up in military, immigration, and educational settings. With the rapid proliferation of mental testing came the need for numerical methods to analyze the resulting data. The correlational methods originally conceptualized by Galton,[1] but elaborated and formalized by Karl Pearson, became the primary method for analyzing mental test data. However, as results of such correlational research accumulated, concerns were raised about the accuracy of correlational methods given that the correlations reported in published literature between tests purported to measure the same or very similar abilities ranged considerably and were even, in some cases, contradictory (Spearman 1904a). Such was the impetus for the development of a body of scholarship dedicated to the technical and mathematical features of mental tests and test data—what would come to be called "psychometric" or "test" theory.[2] It is to these early roots of testing theory this chapter is dedicated.

Before describing in some detail the origins and developments of psychometric theory, there are two distinctions that have been frequently featured in a good deal of contemporary testing theory, which are now most often given only hand-waving acknowledgement, but which, in fact, are integral to understanding the nuances of CVT and all other contemporary validity frameworks. The first distinction is between *classical test theory* and *modern test theory*, the second between *reliability* and *validity*. Each is described briefly in the following section.

Classical Versus Modern Test Theories

There are a great many theoretical results, concepts, and methods that fall under the broad label of *test theory*. Often a distinction is made between two historic periods of testing theory. The first originated in the early part of the twentieth century in the works of Spearman (e.g., 1904a, 1907, 1910), Brown (e.g., 1910), Kelley (e.g., 1916, 1921), among others, and has come to be known as *classical test theory* (CTT). The second, known as *modern test theory* (MTT) , describes a period of psychometric theory that first appeared around the mid-twentieth century with advances in both item response theory (IRT) and factor analytic methods and is based in the works of figures such as Lawley (1940, 1943b, 1944), Tucker (1946), Lazarsfeld (1950), Lord (1952), Lord and Novick (1968), and Birnbaum (1968).

CTT encompasses a set of techniques for describing some fairly basic psychometric properties of test data in terms of the *true score model*, according to which an individual's observed test score is conceptualized as being composed of two non-overlapping parts: a "true score" and a "error" component. Thus, over a population of individuals, the variance of observed test scores is also decomposable into two non-overlapping components, namely, the true score and error variance components, the former representing variability across individuals in the "amount" possessed of the attribute measured by the test, the latter representing variability in the population regarding how well (or poorly) observed test scores represent the individual's "true" score, with low error variance indicating, on average, a relatively more "pure" test of the ability or attribute in question. Although CTT incorporates some item-level analytics, generally the theory is quite narrowly focused on providing estimates of both the degree of measurement error contained within the total test score, as well as of the extent to which the test score predicts (or otherwise correlates with) other variables, the latter taken to be an index of the test's validity.

In contrast, MTT describes a much broader class of theoretical results, most of which presume, however, that the interitem structure of a set of test data may be represented well by one or more latent variable models. Within this framework, observed item-level test data are viewed as “manifestations” or “indicators” of some unobservable attribute (or “latent trait”) (or set of attributes, traits) of interest. Each model specifies the mathematical form of the item/latent trait regressions, and particular implications are drawn and tested on the basis of observed test data. If the data are shown to conform to the model, then an optimal compositing rule may be derived from the model, a composite formed, and an estimate of precision (“reliability”) of the composite calculated. Finally, composite scores demonstrating sufficient measurement precision might be entered into a variety of further analyses of external test score validity (i.e., examining theoretically derived relations with other variables).

Reliability and Validity

There is no more celebrated dyad in psychometric and testing theory than that of reliability and validity. These two psychometric concepts are—at least superficially—treated as signifying the two prerequisites of a “psychometrically sound” test (or, more recently, of scientifically admissible test data or interpretations and/or uses thereof). Testing textbooks are frequently oriented around these two psychometric concepts, and even more general introductory psychology and research methods texts make some reference to the importance of “reliable and valid” measures of focal phenomena.

Although often defined vaguely as the “consistency” of a *measure*, psychometric reliability is broadly defined as a quantitative index of the degree of measurement precision associated with a test (or subtest) score. Technically defined, reliability is the ratio of two variance parameters, namely, of the true score and observed score variances defined under the classical true score model described briefly earlier. The basic idea underlying the latter definition is that a given individual’s score on an item or test will vary over an infinity of hypothetical “in vacuo” replications of a measurement procedure, the expected value (roughly, the arithmetic mean) of this “propensity distribution” being equal to the individual’s true score on the test. From this, the *unreliability* of the observed test score for a given population of individuals and the measurement procedure in question is defined as the proportion of error variance to the total observed score variance. Reliability is its complement; specifically, it is the proportion of true score variance to observed score variance. It constitutes a specific measure of the more general property of the precision of a random variable, more generally. There are different reliability indices associated with different means of obtaining an estimate of reliability (split-half, internal consistency, alternate forms, and test-retest estimates being the most common within typical test evaluation contexts), as well with different types of data (e.g., dichotomous versus continuous item responses; self-report versus rater data, etc.). Reliability is a legacy of CTT, although the term has become generalized to mean the precision or consistency of any test score, regardless of whether the test in question has been analyzed with a classical or modern test theoretic approach. The MTT equivalent to classical reliability is "information," which might apply to either individual items or test scores, and is seen to be a function of the level of ability (trait, etc.) of the test taker and, thus, will vary for a given test (or item) over individual test takers.

Validity is a much broader concept and extends to an array of features of tests and test data, as well as interpretations, uses, and consequences thereof. Historically, psychometric validity was a much narrower concept, and was defined as the population correlation of a test score with any of a number of "criteria" (i.e., theoretically relevant external variables) predicted to be related in specific ways to the attribute purportedly measured by items of the test, the validity of which was in question. Along with advances in testing and psychometric theory, as well as ever-increasing applications of such theory in both research and applied settings, the older, and much narrower, concept splintered into a quite large class of validity concepts, each answering to a relatively specific aspect of the testing enterprise (Newton and Shaw 2014). It is presumed in the current work that any proper analysis of construct validity must acknowledge the quite persistent notion that testing theory—as broad and varied as it has been in scope over the past 100 years or so—has been, minimally, oriented around addressing two primary questions about psychometric instruments and the data resulting from their application: Do they constitute sufficiently precise (or consistent) measurements of *something* (i.e., is the measurement error acceptably low)? And, if so, do they permit the sorts of inferences (and uses) that are desired about the particular attribute of relevance (i.e., does the test data permit *valid* inferences and uses)? The nesting of the second question in the first points to an important and long-recognized feature of the relationship between psychometric reliability and validity: validity requires reliability, but not the converse. That is, to be deemed valid (in one or more senses), test scores must be reliable (i.e., have a high degree of measurement precision), whereas, even an error-free (i.e., perfectly reliable) test score (or inferences/uses based on it considered valid) need not be considered valid if the test score does not measure of the attribute in question.

A more detailed account of the historical, conceptual, and technical developments of test theory—including of reliability, validity, and of the respective contributions of classical and modern test theories—will be taken up in this and the next two chapters. In the present chapter, I begin with a description of the work of Charles Spearman, who provided some of the core foundations of the psychometric theory and practice that would develop throughout the early twentieth century and, as such, were fundamental to the technical and conceptual foundations of CVT.

Spearman's Legacy in Two Important Works

In 1904, Charles Spearman published two papers, both appearing in *American Journal of Psychology*, in back-to-back issues (Spearman 1904a, b), and both involving analysis of the same data (McDonald 1999). The first paper, entitled "The Proof and Measurement of Association between Two Things," is primarily concerned with demonstrating a method for correcting attenuated estimates of correlation indices. In the second paper, "'General Intelligence,' Objectively Determined and Measured" (Spearman 1904b), Spearman describes his theory of intelligence, according to which covariation among different mental measurements are explained by the existence of both a genetically endowed "general intelligence factor" (or "*g*," as it would come to be known) common to all tests, and "specific factors" unique to each individual test. The first of these papers is generally credited as providing the foundations for CTT, and the second for the common factor theory that underlies factor analytic methods, the latter of which anticipate in some important ways modern test theoretic approaches (McDonald 1999).

The Birth of Classical Test Theory

Spearman (1904a, p. 72; emphasis in original) opens the first paper with the claim that although "All knowledge…deals with uniformities," in most cases, knowledge claims are "partial" rather than "absolute." He goes on to say that "In psychology, more perhaps than in any other science, it is hard to find absolutely inflexible coincidences" and that although "there appear uniformities sufficiently regular to be treated as laws…infinitely the greater part of the observations hitherto recorded concern only more or less pronounced *tendencies* of one event or attribute to accompany another." Spearman then questions how it is that after several decades of "laborious series of experiments" the psychologist's knowledge of the correspondence between two things "has not advanced beyond that of laypersons." He dedicates the remainder of the article to attempting to "remedy this scientific correlation."

Spearman organized the paper into two major parts. The first part concerns a description of the principles of correlation and the problem of "accidental deviations," the latter of which he does not define explicitly but describes in terms of "probable error," or variation due to inaccuracies of measurement. Spearman then explicates the "standard" methods of correlation, including the "product moments" methods discovered by Bravais and elaborated by Pearson. He also describes in some detail the advantages and disadvantages of the "rank method" of correlation as well as several "auxiliary" correlational methods, for use when either the Pearson or rank methods cannot be reasonably employed.

The second part of the paper begins with a description of "systematic deviation," which Spearman (p. 76; emphasis in original) contrasts with the "accidental" inaccuracies that are the result of probable error. Whereas accidental errors will "eventually more or less *completely* *compensate one another*," systematic errors, which vary in nature, are "constant," or "non-compensating" inaccuracies. Moreover, accidental deviations might either augment or diminish the correlation but will ultimately "perfectly counterbalance one another," but systematic deviation will always have an attenuating effect on the correlation. To ground this idea, Spearman considers a scenario in which one wishes to ascertain the correspondence between a series of values p, and another series of values q. Due to systematic deviations, only approximations, p' and q', can be observed of the "true objective values," p and q; that is, whereas p' and q' are laden with systematic error, p and q are not. By consequence, the real correspondence of p and q, as measured by r_{pq} , will be attenuated into $r_{p'q'}$, that is, the observed correlation between the approximations p' and q'.

Spearman then spends much of the remainder of the paper demonstrating the amount of attenuation that will occur under varying conditions and presenting corresponding correction formulae that can be applied under these varying conditions in order to "discover the true correlation" between p and q (i.e., r_{pq}) from two or more independent observations of each. The first of the attenuation correction formulae presented by Spearman in the paper is the now familiar expression,

$$r_{pq} = \frac{r_{p'q'}}{\sqrt{r_{p'p'} r_{q'q'}}} \tag{2.1}$$

in which $r_{p'q'}$ is the observed average correlation between the individual measures of p with the individual measures of q, $r_{p'p'}$ is the average correlation between one and another of several independent measures of p, and $r_{q'q'}$ is the same for q. Notably, Spearman did acknowledge the practical difficulty of obtaining two or more observed measures of p and q that are "sufficiently independent" of one another.[3]

Spearman ends the paper with an illustration of his methods of correction for attenuation using correlational results from Pearson's investigations of "collateral heredity." Spearman shows that, when corrected, is it likely that the observed average correlations are underestimated. However, he further claims that, given mental measurements are likely to be affected to a much larger extent than physical measurements by sources of error, “it is difficult to avoid the conclusion that the remarkable coincidence announced between physical and mental heredity can hardly be more than mere accidental coincidence” (p. 98). The point of illustrating such work, Spearman contends, is only to impress upon psychological workers the importance of improving the existing “methodics” of correlational work by introducing correctives such as the correction formulae he presents in his own paper.

Although Spearman’s first paper would seem on the surface to deal mostly with proposing a method for correcting attenuation in correlational indices, there are several other notable implications that might be drawn from his presentation of the problem. The first is the portrayal of observed mental measurements as, at best, “approximations” to “true” or “real” “objective values” and the related notion that observed correlations between two such measurements depart to greater or lesser extents from the “true correlation” between the values measured. In other words, a starting point for Spearman is that mental attributes exist in some objective realm, free from error, and that observed measurements will contain some degree of perturbation, due to either “accidental” or “systematic,” or both, types of deviation from the pure qualities that underlie them. A most deleterious consequence of this “deficiency” is that observed correlation between two measured attributes may dramatically misrepresent the “real” correlation, which if left unaddressed could seriously undermine efforts to establish psychology as suitably rigorous science. A second implication, following from the first, is Spearman’s promotion of mathematical correlation as a methodological foundation for a scientific psychology. Although Spearman was hardly unique at the time in his advocacy of quantitative approaches to psychological inquiry, the legacy of his privileging mathematical correlation as the foundation for psychometric theory and practice is still very much with us today. A third, perhaps less obvious, implication is that the paper includes the earliest articulations of the concept of ‘reliability’ (although Spearman would not use this language until a later paper, Spearman [1910]) and draws attention to the need to develop a special set of statistical theory and techniques for addressing the problem of measurement imprecision. A good deal of the test theory that followed Spearman’s first 1904 paper would be concerned with developing indices for estimating the reliability of mental test scores of varying types. A brief description of these early developments in test theory is provided toward the end of this chapter.

The Birth of Factor Theory: A Prelude to Modern Test Theory (MTT)

In his second 1904 work, "'General Intelligence,' Objectively Determined and Measured" (Spearman 1904b, p. 268), Spearman attempts to bring the correlational methods described in the first 1904 paper to bear on his ideas regarding intelligence and the existence of "General Intelligence" (*g*) in relation to its correspondence to "General Discrimination," defined respectively as the "common and essential" elements of the various forms of the "Intelligences" (such as manifest school examinations and teacher assessments) and "Sensory Functions" (such as discrimination of sound, light, weight, etc.). Spearman takes a rather long route through five chapters (and about 85 pages) to get from the correlational methods he proposed in the earlier 1904 paper to a discussion of the general (common) and specific functions underlying measured intelligence. The latter would become the basis of the "two-factor" theory upon which much of his subsequent work would be founded.

The beginning of the paper recapitulates Spearman's concerns about the methodological weaknesses of experimental psychology and his advocacy of a "Correlational Psychology" as the only feasible remedy for the inconsistencies in (and, in some cases, even contradictions among) experimental findings up to that point in scientific psychology. After describing in detail the history of previous correlational experimental research in psychology, Spearman again diagnoses the cause of this undesirable state of affairs as failure to invoke precise quantitative expression of associations among mental attributes and then subsequently properly account for both accidental and systematic inaccuracies in the measurement of such attributes.

Spearman describes results from a series of his own experiments to illustrate the utility of his two main correction formulae for eliminating the effects of observational errors and irrelevant factors, and thereby "deduce" the "true" correlations that are "of real scientific significance" (p. 256). Finally, he turns to an inquiry into "that cardinal function which we can provisionally term 'General Intelligence'" (p. 205) and its relation to sensory discrimination that was the focus at the time of much of the psychological laboratory work of which Spearman was so critical. Spearman applies his correction methods first to a variety of correlations between specific measurements of sensory discrimination (e.g., pitch) and intelligence (e.g., "School Cleverness"), concluding: "Whenever we have succeeded in obtaining a fairly pure correlation between Sensory Discrimination and Life Intelligence, we have found it amounts to a very considerable value" (p. 268). Spearman then examines the correspondence between averages of specific measures of sensory discrimination and intelligence, respectively, on the basis of which he "arrive[s] at the remarkable result that the *common and essential element in the Intelligences wholly coincides with the common and essential element in the Sensory Functions*" (p. 269; emphasis in original). He summarizes his conclusions regarding these results as follows:

> On the whole, then, we reach the profoundly important conclusion that *there really exists a something that we may provisionally term "General Sensory Discrimination" and similarly a "General Intelligence," and further that the functional correspondence between these two is not appreciably less than absolute.* (p. 272; emphasis in original)

Spearman makes two additional comments that are germane to the present discussion. First, he notes there is a "hierarchy" among individual measures of intelligence; that is, specific measures of intelligence are "variously saturated" with the common intellectual function, with some having higher true correlations with it than others. The evidence he cites is given in a table of correlations (or, what would later be called factor "loadings") between the General Intelligence factor and "specific factors" (e.g., school scores on different subjects, assessments of common sense and cleverness, etc.). The second important comment, implied by the first, is that individual measures of intelligence may be characterized as having a "common" and a "specific" part and, thus, correlations among these measures can be accounted for by the presence of two different *kinds* of factors: a single factor, *g*, common to all intelligence measures and specific factors that are unique to each individual measure of intelligence. Spearman's great insight was that, if the influence of the latter could be controlled by experimental or statistical control, the presence of the common factor could be detected in correlation patterns in test data. Thus, Spearman's "two-factor" theory was born (Bartholomew 1995; McDonald 1999). Spearman would spend the next 40 years of his working life elaborating (and defending) his two-factor theory and developing the early technical foundations of factor analysis, much of this work culminating in his (1927) book, *Abilities of Man* and its "continuation," the posthumously published *Human Ability* (Spearman and Jones 1950).

As with his first 1904 paper, the "General Intelligence" paper recapitulates the importance of recognizing and addressing the deleterious effects of measurement error on correlational indices. However, Spearman goes further in the second paper than merely reaffirming that observed correlations among mental measurements need to be corrected for measurement error. For Spearman, it was the linking of individual measures of intelligence to a common intellectual function that presented the greatest value for the science of psychology at the time. In essence, Spearman proposed the first latent variable model, in introducing an unmeasured variable through observed relations among measured variables (Bartholomew 1995). In doing so, he set into motion a longstanding tradition in testing analysis of using a statistical modeling approach to investigating what are presumed to be the real, but unobservable, attributes that test items measure in common. Although it would be a number of decades after the publication of Spearman's "General Intelligence" before the early developers of MTT would propose methods for modeling the structural relations among item responses and "underlying" (or, "latent") attributes, in many respects Spearman's two-factor model, and his theory that all individual measures of intelligence are underlain by a common function, anticipated such developments.

Major Implications of Spearman's Works

Despite an apparent difference in focus, Spearman's two works were connected in important ways that would have joint implications for developments in validity theory and practice in the early to mid-twentieth century. First, somewhat trivially, the appearance of these two works in the same year by the same author would bond test theory to factor theory (Blinkhorn 1997; McDonald 1999). Much more importantly, they are united by a number of significant conceptual and technical foundations, the linkages among some of which have gone unrecognized on a broad scale. Second, both works are founded on the notion that mental measurements are "impure" or "indirect" reflections of more objective qualities, the latter of which are the true target of psychological scientists. Third, both works advocate rigorously applied correlational methods as an appropriate means of revealing such objective qualities. It is in this respect, perhaps more than in any other, that concerns regarding measurement error (i.e., unreliability) become integrally connected to concerns regarding the validity of measurements. Specifically, since correlations among measures were deemed the essential indicator of the extent to which different measures reflect a common ability, attenuation of correlation due to unreliability became the primary threat to establishing validity of a set of measures *as measures* of that common ability. In other words, reliability of measurements, although not sufficient for establishing validity, became recognized as a necessary condition for validity.

Although it would certainly be overstating it to say that absent Spearman's two earliest works test theory would not have developed, clearly these two works were critical to how early test theory did actually develop, with the imprints of Spearman's works clearly visible in both the classical and modern test theory frameworks. In the following section, a brief overview of the key developments in each of these traditions is given. This is followed by a description of an important change in the tides with respect to the conception of validity as the classical framework began to give way to MTT approaches toward the mid-twentieth century.

Early Developments in Test Theory

Early Classical Test Theory: Emphasis on Reliability of Measurement

In the decades following Spearman's 1904 article, much of the focus of test theory would turn to the development of techniques for estimating reliability of mental test scores. In 1907, in response to criticism (most pointedly from Karl Pearson) that his mathematical results had yet to be substantiated, Spearman provided proofs for two main attenuation correction formulae presented in the first 1904 paper, namely, the formulae for eliminating the effects of "irrelevant factors" and "inaccurate observation" (Spearman 1907). His derivation of the latter formula would include perhaps the first formal (i.e., mathematical) statement of the relationship between observed test variables and the underlying ability of which they are presumed to measure (Levy 1995). This is important because it foreshadows a later emphasis in testing theory on the "latent structure" of test items, and, thus, indicates an early connection between reliability (or measurement precision, more generally) and validity of test scores.

In 1910, Spearman responded to further criticisms that his attenuation formulae, although appropriate for correcting "accidental" deviations, might not be equipped to handle "discrepancies between successive measurements," which cannot be boiled down to "accidental" deviation (Spearman 1910, p. 272). In response, Spearman emphasized that such a "systematic deviation" could be handled through experimental control, but that the remaining "accidental" deviation would still require statistical correction. He suggested a new correction formula, based on a method of dividing the series of measurements for each of the two true values whose correlation is of interest into p and q groups respectively. Taking $p = q = 2$, this correction formal was the first expression of the "split-half" reliability coefficient. In the same paper, Spearman introduced the expression "reliability coefficient" to describe the "coefficient between one half and the other half of several measurements of the same thing" (p. 281) and a formula from which estimates of the reliability of composites of the full set of measures of one or the other attribute (i.e., averages of p or q) could be obtained. Spearman illustrated this formula for the special case in which $q = 1$, in which the formula expresses reliability as an increasing function of test length (i.e., number of measures, or items, that comprise the test).

In an article adjacent to Spearman's, William Brown defined a coefficient measuring the extent to which "the amalgamated results of… two tests would correlate with a similar amalgamated series of two other applications of the same test" (Brown 1910, p. 299), which is equivalent to Spearman's formula, but for p tests. Spearman's and Brown's independently derived formulae would come to be known as the *Spearman-Brown prophecy* (S-B) formula, an index of the effect of test length on the reliability of composites scores, and remains in use today. However, whereas Spearman proposed correlating average measures from two halves of the test to get an estimate of the reliability of the individual measures (an argument in the S-B formula), Brown's method involved correlating two administrations (about 2 weeks apart) of the same series of measures. The two different methods would constitute early definitions, respectively, of "split-half" and "test-retest" estimates of reliability. Such work would be important to establishing methods of producing so-called parallel tests such that the reliability of scores from one or the other could be estimated.

Regardless of the differences in their conceptions of and proposed approaches for estimating reliability, the 1910 works by Spearman and Brown underscore three important results for CTT: (1) an estimate of the reliability of a test score could be obtained by correlating scores of that test with an equivalent test of the same attribute; (2) the reliability of a test score is an increasing function of the number of items included on a test; and, thus, (3) it is possible to determine how many items must be added to a test to obtain a desired degree of reliability for a test score composed of such measures.

In the decades that followed, attempts were made to refine methods for estimating the reliability of test scores. Abelson (1911) and Kelley (1916, 1921, 1924) both provided quantitative expressions of the relationship between individual test scores and "true" scores, the latter defined as the average of the infinity of similar such measures of the same attribute. This echoed Spearman and Brown's emphasis on the importance of developing tests with enough, and suitably comparable, items to ensure sufficiently reliable test scores. Efforts were also made to develop methods for estimating reliability from a single application (or form) of a test in order to circumvent challenges inherent to producing two (or more) equivalent forms of a test, or suitably similar testing conditions for two consecutive administrations of the same test (Cronbach 1951; Guttman 1945; Kuder and Richardson 1937). Such "internal consistency" reliability estimates had not only the appeal of being methodologically efficient, they shed light, once again, on the importance of evaluating the "structure" of a set of item responses for a test purportedly designed to measure a particular single attribute, and, in so doing, reaffirmed the tight bond between reliability and validity.

Axioms of Classical Test Theory

In his text *Statistical Method* (1923), Kelley included a chapter on functions involving correlated measures in which some basic results were presented on the reliability of measurement in terms of regression of true scores on fallible test scores. Four years later, Kelley (1927) published *Interpretations of Educational Measurements,* which was more exclusively dedicated to presenting statistical results and methods relevant to mental measurement. These works were among the first attempts to formalize CTT and the true score model on which it was founded. In 1931, Thurstone published *The Reliability and Validity of Tests* (Thurstone 1931a), in which he expanded on Kelley's treatment, including, among other things, additional sections on different methods for determining the reliability of a test, the effect of test length on validity, and the relations between reliability and validity. This work is the first of its kind to include a relatively comprehensive summary of the then 30-year history of test theory. In (1950), Gulliksen provided a formal summary of the first half-century of test theory in his book, *Theory of Mental Tests*. To say this work provides a thorough summary of mental testing theory up to that point in time would be a gross understatement. It provides a comprehensive account of the first 50 years of technical developments pertaining to psychological testing, and includes derivations of the basic formulas of the classical true score model. Gulliksen's work was also the first to constitute an exhaustive treatment of issues relevant to both constructors and users of psychological tests.

Although the roots of MTT would begin to germinate by late 1940s, CTT reigned as the dominant framework for test theory into the latter half of the twentieth century. Later axiomatic treatments of true score theory were provided by Lord (1959) and Novick (1966), these individual efforts laying the groundwork for their later, now very well-known, collaboration (with contributions by Allan Birnbaum), *Statistical Theories of Mental Test Scores* (1968). In the first three of five parts of the latter work, Lord and Novick recapitulated in extended form many of the CTT results presented in Gulliksen (1950). In the fourth part, they expanded considerably on the previously received, and decidedly narrow, conception of validity as correlation with a criterion, adopting the then relatively new construct validity conception of test validity. The final part of *Statistical Theories,* contributed by Birnbaum, includes an introduction to and description of latent trait theory and of latent trait models and the utility of these for making inferences about examinees' positions on some latent trait. The inclusion of such topics represents a significant departure of Lord and Novick's treatment of test theory from that previously codified in Gulliksen (1950), and a general shift in focus within test theory literature from a presentation of classical to modern test theory results.

Early Developments in Modern Test Theory: Emphasis on Structure

Despite the fact that Spearman's 1904 "General Intelligence" paper would quickly invite controversy (see Bartholomew 1995; Steiger 1996; Steiger and Schönemann 1978), the basic model and methods presented in the paper provided a foundation for a body of psychometric work dedicated to producing theory and methods for investigating the structural relations among test variables and between test variables and underlying attributes. However, whereas Spearman was committed to his theory of intelligence and, thus, to the idea that all the interesting variability among mental measurements is accounted for by *g*, others, most notably Thurstone, would challenge this view. Thurstone questioned the assumption that a single common factor underlies all cognitive functions. He did not believe so, and, thus, extended and developed Spearman's basic factor methods into multiple factor analysis, a larger class of factor models and factor analytic techniques (Thurstone 1931b, 1935, 1947). This seemingly straightforward extension of Spearman's model not only broadened the conception of intelligence as a set of related but relatively distinct cognitive abilities, it also extended the potential applicability of factor analysis to domains not strictly concerned with intelligence (e.g., personality testing, clinical diagnosis, etc.). R. Cattell, Burt, and Guilford, among others, would also be key players in promoting multiple factor analysis for examining the structure of psychometric instruments. However, prior to the mid-1950s, the classical true score model, with its emphasis on estimating true scores, and maximizing the reliability of test scores, remained the dominant test theory framework.

Toward the 1950s, a quite separate line of psychometric theory was beginning to evolve. It would provide the foundations of a theory of item responding that is variously referred to as *item factor analysis*, *latent trait theory*, and (more recently) *item response theory* (McDonald 1999). Because factor models assume linear relations between item responses and factors, and thus, are applicable for tests composed of items having continuous (or pseudo-continuous) response scales, test theory scholars began to recognize the need for psychometric theory and techniques appropriate for tests composed of binary items. The early articulations of this theory may be summarized in terms of a number of key themes.

First, as noted, early latent trait theorists recognized that individual examinees' responses vary as a function of both features of the individual test taker and features of the test, most notably, the form of the item response. To accommodate this, Lawley (1943a) proposed a model within which the probability of an examinee passing a binary item is a function of both the individual examinee's ability level (or, location on the latent trait continuum) and features (parameters) of the individual item.

A second, and related, theme emphasized the precise mathematical form of the relationship between the latent trait and observed item responses. Latent trait theory was developed to model responding to binary items ("pass"/"fail," "correct"/"incorrect"), which cannot be adequately described by linear regressions of item responses on the latent trait because linear regressions are unbounded and, so, permit illogical probabilities of passing an item that fall below zero and exceed 1. Therefore, early trait theorists modeled item-trait regressions[4] in terms of S-shaped functions, bounded below by zero and above by 1, such as the normal ogive (Lord 1952, 1953; Tucker 1946) and logit (Birnbaum 1968, 1969) functions. Such functions imply relatively lower/higher probabilities of response for individuals at lower/higher locations on the trait dimension in comparison to the midrange. As noted, item response models also formalized definitions of item parameters as features of item-trait regressions, namely, item "difficulty" and "discriminating power." Importantly, unlike their analogues in classical item analysis, latent trait models do not assume that item parameters are invariant across populations of examinees that differ in ability, implying a more finessed conceptualization of the relation between test variables and the attributes they are presumed to measure than that described within CTT.

A third theme is the emphasis placed on the "latent structure" of multiple item tests. In a chapter explicating the logical and mathematical foundations of latent structure analysis, Lazarsfeld (1950) introduced the concepts of 'manifest' and 'latent' to describe, respectively, the observed response patterns to test items and the underlying (latent) trait continuum about which inferences are drawn on the basis of the observed responses. Thus, although the language differed, like factor theory, early latent trait theory also presumed that observed associations among the responses to test items were indicative of the influence of a common trait (or "ability," as it was often referred to at the time). However, within latent trait theory, the weaker condition of conditional association in factor theory (i.e., that the correlation between two measures of a common factor disappears when conditioned on the common factor) would be replaced in trait theory by the stronger condition of *local independence*,[5] according to which the multivariate distribution of an entire set of item responses from a test, when conditioned on any fixed position on the latent trait, is the product of distributions of individual item responses. Although local independence refers to the more general property of statistical independence, it became a defining feature of all latent variable models and a fundamental aspect of modern test theoretic approaches.

A fourth theme concerns precision of measurement. Whereas classical test theoretic accounts merely take score reliability to be invariant over different ability levels, latent trait theory defines precision of measurement (a more general concept than reliability) conditionally, specifically, as a function of the level of ability. In other words, as with item parameters, the precision of measurements on a test item (or score) is taken to vary across the trait continuum, generally being higher toward the extremes (Lord 1953). Birnbaum (1968) provided a formal definition of precision of the item response (and item response composites) in terms of item (and test) "information" functions, which give, roughly, a quantity that is inversely proportional to the width of the confidence interval of an estimate of a given examinee's ability (Hambleton and Cook 1977).

There would be tremendous development of both factor and item response theory in the latter half of the twentieth century, in particular as advances in computing capabilities enabled test theorists to apply innovative theoretical results to data. Some of these are described in Chap. 4. For the time being, it is important to note that factor analysis and item response theory, although having somewhat divergent developmental trajectories, would become united under the banner of "modern test theory," a fundamental feature of which is the application of a broad class of latent variable models in the development and evaluation of a wide variety of psychometric instruments.

Clearly, the description of early test theory given here is a mere sketch of a vast body of work that includes many players, most of whom have gone unmentioned. The intention has not been to provide a comprehensive history of the origins and early advances in testing theory. Instead, the aim has been to describe broadly its major contours in order to help situate the theoretical and methodological developments relevant to test validity, and construct validity theory and practice, most particularly, the latter of which will be taken up in the next three chapters.

However, before leaving the present chapter, a few brief words on the status of validity in early testing theory are required. In my admittedly cursory descriptions of early testing, very little has been said about validity. The final section this chapter provides a high level summary of where validity fits into early testing theory.

Early Conceptions of Test Validity

As has been mentioned, CTT was primarily concerned with the estimation of true scores and, thus, with determining methods for maximizing the reliability of observed measurements. Whereas reliability indices quantify how precisely, or consistently, a test score measures *something*, they do not in and of themselves say anything about that something, such as *what* it is and *whether* the items of a test measure it. This is where validity enters the scene.

Newton and Shaw (2014) characterize the early history of validity (i.e., pre-1952) in terms of two major periods: the "gestational period" (from the late 1800s to 1920) and the "period of crystallization" (from 1921 to the early 1950s). The most significant aspect of the former was the rapid and massive growth of the testing movement itself, bringing with it a need for standardized procedures for judging the quality of tests. However, at this same time, there was growing discontent with traditional school achievement exams and some testing scholars began to draw distinctions between different kinds of tests (e.g., between linguistic tests and performance tests, between individual and group tests, between examinations and standardized tests), as well as between professional and scientific testing contexts. Despite the differences among the types of tests and contexts of testing, validity became increasingly recognized as an important property of tests, and toward the second decade of the twentieth century, more and more references to validity began to appear in the literatures of psychology and related disciplines (Newton and Shaw 2014).

However, the concept of 'validity' and methods for assessing it would, according to Newton and Shaw (2014), not become "crystallized" until after 1920. In 1921, what is considered by some to be the first formal definition of validity appeared in a report of the Standardization Committee of the National Association of Directors of Educational Research (NADER, the American Educational Research Association [AERA] in its "embryonic" form; Michell 2009; Newton and Shaw 2014). In this report, the committee contended,

> [t]wo of the most important problems in measurement are those connected with the determination of what a test measures and of how consistently it measures. The first should be called the problem of validity, the second, the problem of reliability.
>
> Members are urged to devise and publish means of determining the relation between the scores made in a test and other measures of the same ability; in other words, to try to solve the problem of determining the validity of a test. (Buckingham et al. 1921, p. 80)

In an article published in the same year, Buckingham defined validity as "the extent to which [tests] measure what they purport to measure" (1921, p. 274). Ruch (1924) and Kelley (1927) would give similar accounts, and this general definition of validity became codified for the next several decades in the testing literature. (In fact, we continue to see variations on this definition in even quite recent accounts (e.g., Anastasi 1982; Colman 2006)).

Given this conceptualization of validity, it is not surprising that the two main approaches to establishing the validity of tests in early testing theory were logical analysis of test content and empirical evidence of correlation (Newton and Shaw 2014). A common approach was to use content analysis to establish a logical link between a test (e.g., of academic achievement) and a criterion (e.g., teacher assessments), and then validate the test against the criterion (Kane 2016). Likely in no small part due to Spearman's promotion of Pearson's correlational methods for estimating true scores from observed scores, the latter, empirical approaches to validation would come to dominate. In particular, since test validity was formulated in terms of tests measuring what they are purported to measure, correlating test scores with that which they are alleged to measure followed logically. This gave rise to the notion of the "criterion" of a test. In the idealized case, the criterion is the outcome the test is intended to measure and, therefore, it was thought that a valid test is one that predicts well this outcome. Thus, in the early decades of testing theory and practice, validating a test was predominately a matter of calculating the "validity coefficient" of a test by correlating scores from the test with some criterion deemed suitably representative of whatever the test was believed to measure. This conception of and approach to evaluating validity would dominate under CTT (cf. Thurstone 1931a).

Despite its prominence prior to (and even beyond) the mid-twentieth century, aspects of the classical account of validity would quickly come under scrutiny. Among the issues identified, perhaps the most persistent concerned how to ascertain a suitable criterion and whether it is appropriate to assume all tests lend themselves to the identification of a single clear criterion. In fact, given the elegant simplicity of the validity coefficient, multiple criteria were often employed, giving rise to the notion that a test is valid for anything with which it correlates (Sireci 2009). Because the scores of particular tests often correlated equally well with different criterion measures, the authority of criterion validity was increasingly called into question. Moreover, test theory scholars began to recognize that not only test scores, but criteria often have at least some degree of measurement error and that this could distort, quite seriously in some cases, assessments of (criterion) validity (Jenkins 1946).

By the 1940s, testing scholars were beginning to recognize that a one-size-fits-all conceptualization of validity and approach to validation would not suffice and called for a more comprehensive treatment (Sireci 2009). Validity became differentiated into different types, one of the most fundamental distinctions being that between "logical" (i.e., based on logical analysis of test content) and "empirical" (i.e., based on correlational evidence) (Cronbach 1949; Rulon 1946). Test theory scholars also began to eschew the "of the test" language used in relation to validity, recognizing that validity depends on the purpose to which a test is put, rather than residing in the test itself. Importantly, developments in factor analysis and latent trait theory would substantially contribute to a broadening of validity to include "factorial validity" and consideration of procedures for assessing it. Guilford (1946, 1954) was a particularly vocal proponent of using factor analysis to define validity and validate tests (Sireci 2009).[6]

Although practical approaches to assessing validity remained faithful to classical conceptions well into the mid-twentieth century, a growing emphasis on the structure of test variables led to a shift in focus from validity in terms of correlations of test scores with criteria, to validity in terms of structure (among other things). This new emphasis on structure generated further discussion among test theory scholars as to how validity should be conceptualized and assessed and, ultimately, opened the door for change in the conceptions of testing validity held and approaches adopted for validation. This tide change in validity theory provided fertile ground for the development of CVT.

The following chapter opens with a description of how the 'validity' concept began to splinter in the late 1940s and early 1950s, and how this would lead to a call for the establishment of standards for how testing validity ought to be conceptualized and approached. Debates that arose in the 1940 s around the roles of, and distinctions between, intervening variables and hypothetical constructs are also summarized. This is followed by a description of a number of key works, precursors to the first formal articulations of construct validity in the Technical Recommendations and C&M, each of which would show their imprint on these two foundational documents of CVT.

Notes

1. Although Galton is credited with defining the statistical concept of 'correlation,' the mathematical foundations of correlational methods are acknowledged as residing in the work of astronomer, Auguste Bravais, but in relation to error theory rather than for the explicit purpose of providing a mathematics of association (Denis 2001).
2. However, as will be illustrated, from its inception, test theory—although concerned with various mathematical features of measurements—would be interwoven throughout with substantive psychological theory, in particular with theories concerned with the heritability of intelligence and other psychological traits.

3. This difficulty would later lead to the development of different methods for producing such independent measures (e.g., split-half, alternate forms, test-retest), methods often taken, mistakenly, to be different "types" of reliability.
4. Tucker (1946) referred to these as "item characteristic curves," which became the convention in latent variable theory. However, Larzarsfeld (1950) used the term 'traceline,' which he would also define as the product of the individual item tracelines for the probability of passing all the items on the test. Lord (1952) also specified "test characteristic curves," but in terms of a curvilinear regression of the total test score (typically an unweighted sum of the binary items) on the latent trait.
5. Lord (1953) called this "homogeneity."
6. Of course, Spearman's two-factor theory was in many respects one of the earliest examples of this sort of approach to getting at the validity of mental measurements, but was narrowly confined to a consideration of how various ways of measuring intellectual ability can be seen to manifest a unitary function, *g*.

References

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

Anastasi, A. (1982). *Psychological testing*. New York: Macmillan.

Bartholomew, D. J. (1995). Spearman and the origin and development of factor analysis. *British Journal of Mathematical and Statistical Psychology, 48,* 211–220.

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

Birnbaum, A. (1969). Statistical theory for logistic mental test models a prior distribution ability. *Journal of Mathematical Psychology, 6,* 258–276.

Blinkhorn, S. F. (1997). Past imperfect, future conditional: Fifty years of test theory. *British Journal of Mathematical and Statistical Psychology, 50,* 175–186.

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

Buckingham, B. R. (1921). Intelligence and its measurement: A symposium. XIV. *Journal of Educational Psychology, 12,* 271–275.

Buckingham, B. R., McCall, W. A., Otis, A. S., Rugg, H. O., Trabue, M. R., & Courtis, S. A. (1921). Report of the standardization committee. *Journal of Educational Research, 4,* 78–80.

Cattell, J. M. (1890). Mental tests and measurements. *Mind, 15,* 347–380.

Colman, A. M. (2006). *A dictionary of psychology*. Oxford: Oxford University Press.

Cronbach, L. L. (1949). *Essentials of psychological testing*. New York: Harper & Brothers.

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

Denis, D. J. (2001). The origins of correlation and regression: Francis Galton or Auguste Bravais and the error theorists? *History and Philosophy of Psychology Bulletin, 13*(2), 36–44.

Gregory, R. J. (2004). *Psychological testing: History, principles, and applications* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6,* 427–439.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

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

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

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

Jenkins, J. G. (1946). Validity for what? *Journal of Consulting Psychology, 10,* 93–98.

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23,* 198–211.

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

Kelley, T. L. (1921). The reliability of test scores. *Journal of Educational Research, 3,* 370–379.

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

Kelley, T. L. (1924). Note on the reliability of a test: A reply to Dr. Crum's criticism. *The Journal of Educational Psychology, 14,* 193–204.

Kelley, T. L. (1927). *Interpretation of educational measurements*. Yonkers-on-Hudson, NY: World Book.

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

Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh, A, 60,* 64–82.

Lawley, D. N. (1943a). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, 61,* 273–287.

Lawley, D. N. (1943b). The application of the maximum likelihood method to factor analysis. *British Journal of Psychology, 33,* 172–175.

Lawley, D. N. (1944). The factorial analysis of multiple test items. *Proceedings of the Royal Society of Edinburgh, 62 A,* 74–82.

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. Clausman (Eds.), *Measurement and prediction* (pp. 362–412). Princeton: Princeton University Press.

Levy, P. (1995). Charles Spearman's contributions to test theory. *British Journal of Mathematical and Statistical Psychology, 48,* 221–235.

Lord, F. (1952). A theory of test scores. *(Psychometric Monographs No. 7*). Richmond, VA:Psychometric Corporation. Retrieved from http://www.psychometrika.org/journal/online/MN07.pdf.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517–549.

Lord, F. M. (1959). An approach to mental test theory. *Psychometrika, 24,* 283–302.

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

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Michell, J. (2009). Invalidity in validity. In R. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 11–133). Charlotte, NC: Information Age Publishing.

Newton, P. E., & Shaw, S. (2014). *Validity in educational and psychological assessment*. London: Sage.

Novick, M. R. (1966). The axioms and principle results of classical test theory. *Journal of Mathematical Psychology, 3,* 1–18.

Ruch, G. M. (1924). *The improvement of the written examination*. Chicago, IL: Scott, Foresman.

Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review, 16,* 290–296.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19–37). Charlotte, NC: Information Age Publishing.

Spearman, C. (1904a). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101.

Spearman, C. (1904b). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15,* 201–292.

Spearman, C. (1907). Demonstration of formula for true measurement of correlation. *American Journal of Psychology, 18,* 160–169.

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

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. London: MacMillan.

Spearman, C., & Jones, L. L. W. (1950). *Human ability*. London: MacMillan.

Steiger, J. H. (1996). Coming full circle in the history of factor indeterminacy. *Multivariate Behavioral Research, 31,* 617–630.

Steiger, J. H., & Schönemann, P. H. (1978). A history of factor indeterminacy. In S. Shye (Ed.), *Theory construction and data analysis in the behavioral sciences* (pp. 136–178). San Francisco: Jossey-Bass.

Thurstone, L. L. (1931a). *The reliability and validity of tests*. Ann Arbor, MI: Edwards Brothers.

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

Thurstone, L. L. (1935). *The vectors of mind*. Chicago: University of Chicago Press.

Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago: University of Chicago Press.

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

Young, K. (1923). The history of mental testing. *The Pedagogical Seminary, 31,* 1–50.

http://www.springer.com/978-1-137-38522-2

Validating Psychological Constructs
Historical, Philosophical, and Practical Dimensions
Slaney, K.
2017, XI, 308 p. 1 illus., Hardcover
ISBN: 978-1-137-38522-2
A product of Palgrave Macmillan UK