VALIDITY in EDUCATIONAL & PSYCHOLOGICAL ASSESSMENT

PAUL E. NEWTON & STUART D. SHAW

CAMBRIDGE ASSESSMENT

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

Established over 150 years ago, Cambridge Assessment operates and manages the University's three exam boards and carries out leading-edge academic and operational research on assessment in education. It is a not-for-profit organisation developing and delivering educational assessment to eight million candidates in 170 countries every year.

UNIVERSITY *of* CAMBRIDGE
Local Examinations Syndicate

Cambridge Assessment is the brand name of the University of Cambridge Local Examinations Syndicate, a department of the University of Cambridge. Cambridge Assessment is a not-for-profit organisation.

CHAPTER 2

THE GENESIS OF VALIDITY: MID-1800s–1951

This chapter explores the period that we describe as the 'early years' as far as validity is concerned, which can be divided into a gestational period from the mid-1800s–1920, and a period of crystallization from 1921–1951.

A gestational period (pre-1921)

The emergence of validity as a formal concept of educational and psychological measurement can only be understood in the context of major developments in testing for educational, clinical, occupational and scientific purposes which occurred during the second half of the 19th century, and the first half of the 20th century, particularly in England, France, Germany and the USA.

The second half of the 19th century was a period during which school and university entrance examinations blossomed. The University of London was established in 1836 to set and regulate examinations, which included matriculation examinations: that is, entrance examinations for its federated colleges. In 1871, the university announced that matriculation would provide exemption from the entrance examinations of other institutions, including the Royal Military College and the Royal College of Surgeons (Cain, 1982). The establishment of Local Examinations by the universities of Oxford and Cambridge in 1857 and 1858, respectively, paved the way to the improvement of education in secondary schools (Roach, 1971). Similarly, the India Act 1853 led to the establishment of examinations as a mechanism for impartial selection within the Indian and home civil services (Montgomery, 1965; Sutherland, 1984). In the USA, Horace Mann promoted the virtues of the written examination, claiming their superiority over older methods such as the oral quiz on the basis of their impartiality (given that

the same questions were submitted to all scholars of the same grade in all schools), but also on the basis of their reliability, economy and practicality (Ruch, 1929). Influenced by Mann, the first written examinations in the USA were introduced by school examining committees in Massachusetts in 1845, and used to determine high school graduation (Linden and Linden, 1968).

Shortly after his half-cousin Charles Darwin had published *The Origin of Species*, Francis Galton published *Hereditary Genius* (1869). He developed and pioneered the use of new statistical techniques which he applied to the study of heredity, laying a foundation for the study of individual differences as a discipline in its own right. In a commentary appended to Cattell (1890), Galton prefigured what would become known as the empirical approach to validation, proposing that 'the sets of measures should be compared with an independent estimate of the man's powers' to determine which were the most instructive (Galton, 1890: 380). Numerous British psychometric pioneers developed Galton's statistical approach, including Karl Pearson, Charles Spearman, William Brown, Cyril Burt and Godfrey Thomson. Spearman was particularly influential in applying the method of correlation to psychological data (e.g. Spearman, 1904), and his simplification of the technique made it accessible to the run-of-the-mill psychologist and educationist (Burt, 1924).

In France, during 1904, Alfred Binet had been approached by the French Ministry of Public Instruction to serve on a committee tasked with the problem of 'backward children'. With Herbert Simon, he devised a series of tests of increasing difficulty, designed to discriminate between unmotivated and incapable children, for placement of the incapable into special classes for the 'mentally deficient'. The first scale was released in 1905, with revisions published in 1908 and 1911 (Burt, 1924; Sutherland, 1984). The Binet-Simon tests were translated and adapted by enthusiastic psychologists across the world, including: Henry Goddard, Robert Yerkes, Lewis Terman and Arthur Sinton Otis in the USA; Cyril Burt in England; Zaccaria Treves and Umberto Saffiotti in Italy; and William Stern in Germany. Under the guidance of Yerkes, related tests of intelligence were developed for use in the selection of army recruits, and were administered to more than 1.5 million North Americans during the First World War. The success of this initiative was such that group tests were rapidly introduced into many North American universities and schools, typically being administered at entrance or at times of promotion. It also stimulated the growth of vocational testing internationally such that in 1924, Burt observed that most of the civilized countries of the world now possessed institutes of vocational guidance in which trained psychologists carried out vocational tests and offered vocational guidance.

Toward the end of the 19th century, the foundations of the measurement movement were being laid in the USA by James McKeen Cattell, following in the tradition of Galton (Linden and Linden, 1968), as well as by researchers such as Herbert Rice, who wished to bring scientific methods based on the use of achievement tests to the study of education (Scates, 1947). It was Cattell (1890) who introduced the term 'mental test'. Shortly after the turn of the century, Edward Thorndike helped to establish the sub-discipline of educational psychology, with the publication of a textbook describing the development of tests for

the prediction of educational success (Thorndike, 1903) and a textbook on the theory of mental and social measurements (Thorndike, 1904). Following Thorndike's lead, educational researchers including Calvin Stone, Stuart Courtis, Burdette Ross Buckingham, Leonard Ayers, Guy Montrose Whipple and many others turned their hands to the construction of tests designed to support instructional processes. As these tests became increasingly accessible to teachers, their sale and use grew to unprecedented proportions.

Of particular significance to the birth of the concept of validity was growing discontent with the traditional school achievement examination. The seeds of this discontent had been sown in North America by researchers such as Meyer (1908), who demonstrated the unreliability of marks from written examinations, and Starch and Elliott (1912, 1913), who followed in his wake. Many believed that the written examination suffered from a major defect: its results could not be evaluated fairly by human minds. In contrast, new-style tests – based on simple recall, sentence completion, true/false or multiple-choice selection and suchlike – were freed from this 'personal equation' (Ruch, 1929).

Concern over the personal equation focused the minds of educational assessors, psychological investigators and personnel selectors on developing more objective methods of assessment. The measurement movement was really beginning to take root now. Although the movement had an international following, its impact was particularly pronounced in the USA, which might help to explain why validity theory began (and largely remained) a product of North America. The movement fostered an industry committed to the development and publication of standard(ized), tests and to the promotion of new-style objective tests. As Ruch explained in 1929, objective tests were basically standard tests without the refinements of experimental study and standardization. (His ambition was to bring the technology of objective testing to the teaching profession, in order to help overcome what he judged to be a fundamental limitation of standard tests: their lack of alignment with local curricula.) As the movement progressed, often the terms 'standard test' and 'objective test' were used interchangeably.

During the second decade of the 20th century, the publication of standard tests mushroomed. In a collection of reports from the Committee on Standards and Tests of the National Council of Education, Courtis (1916) reported a remarkable growth in interest in the measurement movement, indexed by the expansion in use of his standard research tests (mainly his arithmetic tests) from one school in 1909 to a despatch of 455,007 to instructors in 42 states between August 1914 and August 1915. Not only were tests and testing mushrooming; research into tests and testing expanded similarly. In a review of Whipple (1915), a new edition of the *Manual of Mental and Physical Tests*, Freeman (1917) observed that in the five-year period between its publication and revision, the number of references reporting experimental research into the focal tests had more than doubled, from 190 to 390. Therefore, the period from 1910 to 1920 was a decade of increasing focus on the validity of mental measurement, although not necessarily described as such.

To summarize the state of play by the end of the gestational period, interest in structured assessment had grown exponentially and already had evolved substantially;

and quite distinct sub-domains of educational and psychological measurement had begun to establish. These included:

- *professional communities* – where written examinations and tests were used for clinical diagnosis, assessing educational achievement and occupational guidance and selection; and
- *scientific communities* – where tests were devised to explore the very structure of ability, intelligence and other such personality characteristics, in terms of which individuals were supposed to differ, often on the assumption that such differences might be innate.

In addition to distinctions that were emerging between sub-domains, distinctions had been drawn between different kinds of test: for example, between linguistic tests and performance tests, individual tests and group tests, written examinations and standardized tests. Perhaps most importantly, the correlation coefficient had become very widely recognized, and was beginning to be used as a tool for judging the quality of tests.

A period of crystallization (post-1921)

It is unclear when or how the term 'validity' began to acquire special significance among measurement specialists. Early examples of its use can be identified as far back as the 19th century (see Cattell and Farrand, 1896, in von Mayrhauser, 1992), but it seems likely that it was not until the second decade of the 20th century that the term began to take root in the lexicon of researchers and practitioners. Even then, authors seemed to be using it in different ways. For example, in 1914 Frank Freeman, who since 1911 had contributed an annual report on tests to the *Psychological Bulletin*, referred to 'the technique and validity of test methods' (Freeman, 1914: 253; see Schouwstra, 2000). The following year, Terman et al. evaluated 'the validity of the intelligence test' (1915: 562) and 'the validity of the IQ' (1915: 557). The year after, Starch (1916) referred to 'the validity or the fairness of these measurements' (1915: 3), and Thorndike discussed 'the essentials of a valid scale' (1916: 11).

In 1921, in the USA, the National Association of Directors of Educational Research established and promoted an intention to formalize and standardize both the procedures and the concepts of educational measurement. Their problem was not a lack of concepts and procedures, but a lack of consistency in their application. They were not the first to attempt to bring an element of standardization to testing practice in the USA. The American Psychological Association (APA) had established committees to achieve similar goals twice previously: first in 1895, under Cattell; and second in 1906, under James Rowland Angell. These committees sought to improve standardization through establishing norms and identifying and promoting the 'best' tests and procedures. In 1919, the APA appointed a qualifications committee to explore the potential for professional certification, in

response to concerns that mental tests were being widely used for psychological diagnosis by individuals unqualified to do so. Reflecting on these various attempts at standardization and control, Fernberger (1932) concluded that they had all resulted more or less in total failure. He concluded that the APA saw itself as an organization of scientists, not professionals, and lacked the will, let alone the ability, to control its members.

In a 'Report of the Standardization Committee' of the National Association of Directors of Educational Research, the returns from a questionnaire to members were discussed, which revealed 'a practically unanimous sentiment in favour of the publication of an official list of terms, procedures, etc.' (Buckingham et al., 1921: 78). In response, the committee proposed definitions for terms including 'scale', 'standard scores', 'average', 'achievement', 'capacity' and suchlike. The goal of the committee and membership was for uniformity and certainty of interpretation in relation to these technical terms. Indeed, members were asked explicitly to try to conform to these regulations when preparing material for publication, or else to provide (in a footnote) the reason why they had chosen to use an alternative form. In addition, the committee recommended that makers of tests should be careful to investigate the effect of every factor involved in standardization, specifying the following component operations:

1 preparation and selection of test material
2 experimental organization of the test and instructions for giving the test
3 trial of a tentative test to determine value of elements, gross validity, reliability and optimum conditions of giving, scoring, etc.
4 final organization of the test
5 final formulation of conditions under which the test is to be given, scored, tabulated and interpreted
6 official determination of validity
7 official determination of reliability
8 official determination of norms.

In a subsequent section headed 'Problems', the committee recommended as follows:

> Two of the most important types of 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: 80)

Thus, it would seem, the very first official definition of validity was proposed; and members of the National Association of Directors of Educational Research were challenged to develop and promote new methods of validation. Within a few years

the early formulation had been modified slightly, to become what might be described as the *classic definition* of validity:

> By validity is meant the degree to which a test or examination measures what it purports to measure. (Ruch, 1924: 13)

That concepts for theorizing the evaluation of measurement quality should come to the fore during the 1920s was not incidental. This was a direct response to the rapid expansion of tests and testing in the hands of converts to the new style. Ruch observed that 'an already bewildering situation is daily becoming more aggravated' (1925: 349), referring to the challenge routinely faced by school superintendents, directors of research and so on, of selecting the 'best' tests to use. Unfortunately, often the enthusiasts creating these new tests were disinclined to be critical of their instruments or the interpretation of their results. Thorndike and Hagen characterized the years between 1915 and 1930 as a 'boom' period, during which new tests 'multiplied like rabbits', and '[m]any sins were committed in the name of measurement by uncritical test users' (Thorndike and Hagen, 1969: 6).

Existing accounts of the early years

Although there are plenty of accounts of the history of validity (e.g. Geisinger, 1992; Shepard, 1993; Kane, 2001) descriptions of the early years tend either to be omitted entirely, or presented in a manner that seems a little over-simplistic. In many accounts, the early years hardly figure at all (e.g. Langenfeld and Crocker, 1994; Hubley and Zumbo, 1996; Gray, 1997; Jonson and Plake, 1998). In fact, it is not at all uncommon for historical accounts (e.g. Moss et al., 2006) to give the impression (at least) that all of the important developments are traceable through successive editions of *Educational Measurement* (beginning with Lindquist, 1951) and the *Standards* (beginning with APA et al., 1954). When the true diversity and sophistication of the early years becomes apparent, this impression seems implausible. In many ways, developments in validity theory and validation practice from the middle of the 20th century onwards are simply elaborations of insights that had been established far earlier.

When the early years do figure in historical accounts, they are often a little oversimplified in the sense of focusing on certain, apparently unsophisticated features at the expense of many other, more sophisticated ones. Geisinger (1992) emphasized the relationship between validity and correlation in the period prior to the 1950s and linked this to the idea of 'dust-bowl empiricism': the elevation of empirical evidence at the expense of logical analysis. Shepard (1993) characterized the 1920s to the 1950s as a period of deference to test-criterion correlations, typically judged in terms of a single coefficient. During the 1940s, she observed, the concept of validity came to be synonymous with the predictive correlation coefficient. Similarly, Kane (2001) characterized the early years as the 'criterion' phase, stressing the

dominance of test-criterion correlations, and suggesting that the criterion was basically taken for granted during this period. Cronbach (1971: 443) observed that the theory of prediction was very nearly the whole of validity theory until about 1950, arguing that only recently had researchers relegated the theory of prediction and elevated descriptive and explanatory investigations – a point that was echoed recently by Brennan (2006). Messick (1989a: 18) also claimed that thinking on validity shifted fundamentally from prediction, during the early years, to explanation. Our own research, based on early textbooks and journal articles of the period, has uncovered a more subtle and interesting story. There is undoubtedly a sense in which the 'criterion phase' captures something very important about the early years; yet at the same time it is indeed an over-simplification, and to characterize this period simply as the 'prediction phase' is worse still.

The first official definition of validity – the degree to which a test measures what it purports to (i.e. is supposed to) measure – needed to be operationalized for validation purposes. This required a criterion by which validity could be judged, therefore the idea of a criterion was central to early accounts. Yet the criteria in question were many and varied, and approaches to establishing validity equally diverse and numerous. This was not exclusively a period of 'criterion-oriented validation', as that term came to be understood from the mid-1950s (Cronbach and Meehl, 1955): that is, a period during which only predictive or concurrent approaches to validation were utilized. For example, it is not true that issues of 'content validity' were rarely, if ever, considered during this period, although the term was not invented until much later.

In fact, content considerations were always fundamental to the design and evaluation of all tests (all tests need content), although content considerations were particularly significant for achievement tests. Admittedly, often the logical analysis of content was supplemented by empirical evidence of correlation – and content considerations were not as robust as they ought to have been in certain quarters – but content considerations always came first. Furthermore, although there is certainly an interesting tale to be told about the reconceptualization of content sampling theory under the influence of the measurement movement, and although there were other ways in which the technology of mental testing compromised the content sampling ideal, this does not mean that the principle of content sampling was somehow absent from early debate over the design and evaluation of tests. If reference to the 'criterion' or 'predictive' phase should give the impression that content concerns were universally trivialized during the early years, this is not true, certainly not as far as mainstream educational assessment was concerned. We should recall that the concept of validity was born to, and moulded by, the National Association of Directors of Educational Research.

If the concept of validity was not quite so narrow and unsophisticated during the early years, then why might such giants of validity theory as Cronbach, Messick, Kane and Shepard seem to suggest otherwise? For example, Cronbach not only lived through, but even worked through, a substantial chunk of the early years, studying full-time on his PhD in education from 1938 and subsequently working as an assistant to Ralph Tyler in his landmark Eight-Year Study (Cronbach, 1989b) – the significance

of which will become clear shortly. Messick, too, studied for his PhD toward the end of the early years. We can only speculate as to why the more interesting story has not been conveyed through existing historical accounts, so here are a few speculations.

First, many accounts of the history of validity have been written during the past 20 years or so, to make sense of the transition from a trinitarian conception of validity (in which content, criterion and construct validity stood alongside each other) to a unitarian one (whereby construct validity came to subsume the other two). As mentioned previously, the trinitarian conception had its roots in the very first *Standards* document (APA et al., 1954). As such, 1954 makes an obvious starting point for a history of validity, and successive editions of this publication, alongside successive editions of the text that many measurement specialists think of as the 'holy book' of their field, *Educational Measurement*, certainly mark crucial watersheds. Unfortunately, the omission of any discussion of the early years is counterproductive, since any sense of continuity between classical and current conceptions of validity is lost. For example, the maxim that 'all validity is construct validity' is not too dissimilar from the earliest definition of validity as the degree to which a test measures what it is supposed to measure (Newton, 2012a). In short, 1954 does not mark the beginning of validity theory, but an unhelpful detour from a journey that began on a promising path many decades earlier. Teaching validity from a baseline of 1921 makes far more sense of present-day theory and practice than starting from a baseline of 1954. Only with reference to developments during the early years can the trinitarian characterization of validity and validation be really understood.

Second, there were no seminal reference works on validity theory between 1921 and 1951. More precisely, there were many seminal works, but none which seem to have resulted in early consensus, and therefore none which acted as a true reference point for educational and psychological measurement. This was not true of later years, whereby each new generation of professionals could refer to a revised section on validity within successive editions of the *Standards*, or to a new section on validity or validation within successive editions of *Educational Measurement*.[1] Perhaps there were too many seminal works during the early years for a coherent tradition to emerge, with each new theorist promoting a slightly different perspective. The 1920s was particularly prolific for educational measurement, witnessing the publication of numerous important and influential textbooks (e.g. McCall, 1922; Monroe, 1923; Ruch, 1924; Kelley, 1927; Ruch and Stoddard, 1927; Ruch, 1929). Perhaps more importantly, in addition to differences in perspective between authors working *within* sub-domains, larger differences were beginning to emerge between authors working in *different* sub-domains. Thus, accounts written in the context of aptitude testing within organizations (e.g. Hull, 1928; Bingham, 1937) had a somewhat different emphasis from those written in the context of achievement testing within schools. Accounts written from a more academic perspective, with an emphasis on psychology and psychometrics, again had a somewhat different feel (e.g. Thurstone, 1931; Guilford, 1936). However, this did not reflect the emergence of sharp dividing lines between sub-professions of educational and psychological measurement: for example, directors of education had just as much interest in intelligence tests as they did

in special aptitude tests or school achievement tests. Had types of test clustered more clearly within sub-domains of professional practice, then validity theory might have developed a clearer identity right from the outset – or more likely still, a number of different identities. In short, there were differences of emphasis during the early years but no dominant approach, so it is actually quite hard to characterize the period.

Third, the sub-domains of professional practice tugged validity theory in different ways during different phases, and the 1940s were particularly dominated by the war effort. In fact, both world wars had a major impact on the perception of testing and validation. The large-scale implementation of mental testing during the First World War was facilitated by Otis, who had developed the technique of group testing: the administration of a single test to a large number of people at the same time, in contrast with the traditional approach of administering tests individually, person by person. He also had devised a method of scoring by stencil, which meant that tests could be marked very rapidly. Under the direction of Yerkes, group tests for military aptitude were developed: the Army Alpha and Beta. Their widespread adoption gave the technology of mental testing such enormous publicity and prestige, that following the war there was a mad rush to transfer the methods to other occupational settings. According to Osburn (1933), the First World War not only brought the word 'criterion' into more general use, it also led to the widespread perception of test construction as a process that was far more empirical than logical, as tests often came to be mechanically constructed to optimize the prediction of criterion measures. Taken to an extreme, this approach to test construction came to be much maligned, being described as 'blindly empirical' (Peak, 1953: 288). The adoption of this empirical approach by many test developers during the early years helps to explain why commentators often have emphasized the dominance of 'prediction' theory. Yet, to characterize everything that happened prior to 1952 in this manner would be misleading, since it would tell only one side of a far more complex and fascinating story.

Given the vast literature on test theory and practice from the middle of the 19th century to the middle of the 20th century, and given the range of professional and academic perspectives that it embraces, it might seem impossible to construct a plausible, comprehensive and even moderately objective report on the history of validity during the early years. We fully acknowledge the scale of this challenge. Expressing similar sentiments, Tim Rogers observed that:

> Trying to piece together exactly how validity emerged in the testing enterprise is like an archaeologist [sic] trying to construct a dinosaur from a few bones. Much of the critical information is not readily available and has to be filled in by guesswork. (Rogers, 1995: 245)

The following account is our reconstruction of a long and productive period that we cannot claim to know well, and have glimpsed only partially through the literature that we have managed to track down. We admit that our narrative may be slightly nuanced because we wish to tell a slightly different story from the one that is often told, and want to highlight synergies with our account of subsequent periods.

Explaining the caricature

Before exploring the evidence which is suggestive of greater diversity and sophistication during the early years, it is worth pausing to reflect further why it might have become known as the 'prediction' phase. There are at least two major factors that, in combination, might help to explain this caricature: first, the widespread adoption of blindly empirical methods, particularly for the purpose of aptitude testing; second, a degradation of the classic definition over time, as the preferred *method* for investigating validity came to be mistaken for a *definition* of validity in its own right. The second factor appears to involve a three-stage deterioration from:

1 quality of measurement supported by the test, to
2 degree of correlation between the test and its criterion, to
3 coefficient of correlation between the test and a criterion measure.

The following sections attempt to illustrate how this definitional degradation might have occurred over time, while the adoption of blindly empirical methods will be discussed subsequently in the section 'Validity and special aptitude tests'.

From quality of measurement to degree of correlation

In one of the first textbooks of its kind, *How to Measure in Education*, McCall (1922) reflected on the earliest formal definition of validity. Asking his reader how we might know whether a test measures what it purports to measure, he responded that we can know what it measures 'only by its correlations' (McCall, 1922: 204). McCall identified two methods for determining this correspondence for achievement tests:

1 follow up the testing with prolonged careful observation of how pupils demonstrate the proficiency in everyday life – that is, rank them first on the test, and then in real life, and correlate the ranks
2 test a population of pupils whose proficiency is already known – that is, rank them first on the proficiency, and then on the test, and correlate the ranks.

In both cases, the criterion against which the validity of the test was to be judged was the pupil's actual, real-life proficiency. If the ranking given by the test agreed with the actual, real-life ranking, then the test was valid for measuring that ability. McCall's *degree of correlation* elaboration of the classic definition was essentially a conceptual abstraction. It envisaged a hypothetical, true proficiency rank as the absolute criterion against which, in theory, the actual, observed proficiency rank from the test ought to be judged. In fact, even McCall was prepared to accept that there was probably no such thing as a single true proficiency rank, and he preferred to think in terms of a range of true proficiency ranks across a family of similar real-world situations (McCall, 1922: 209). Yet the abstraction provided a useful heuristic to guide validation.

The degree of correlation conception naturally recommended a particular approach to validation: the careful construction of a comprehensive criterion measure according to which the shorter, more practical test could be judged. Thus, McCall introduced the idea of prolonged careful observation of pupils in real-life situations, to determine their true proficiency as precisely as possible. This more comprehensive judgement of proficiency then could be used as a criterion measure for test validation purposes. Importantly, even this comprehensive judgement was still only a measure of the criterion – a measure of the pupil's actual, real-life proficiency – and not the criterion itself. Ultimately, the criterion was more of a conceptual abstraction than an empirical operationalization, yet the criterion measure was as faithful a representation of the criterion as could be realistically achieved.

During the early years, different approaches to the development of criterion measures were proposed and utilized. Results from tests to be validated came to be correlated against:

- the judgement of suitably qualified experts, especially teachers
- results aggregated across multiple existing tests purporting to measure the same thing
- results from specific tests that commanded particular respect with the passage of time.

In relation to studies such as these, during the 1920s the elaboration of the classic definition of validity still seemed to be conceptual. That is, although correlation was central to the process of validation, the definition of validity and the method of validation were discrete. Importantly, validity was not framed in terms of prediction, in the sense of predicting future proficiencies, but in terms of correlation: that is, the correlation between actual test result and hypothetical true proficiency.

From degree of correlation to coefficient of validity

As time went by, textbook authors restated the degree of correlation definition in their own words, influenced by their own interests and settings. As these proliferated, sometimes the subtlety of the classic definition was lost, as criteria came to be described less as conceptual abstractions, and more as concrete measures. For example:

> The *validity* of a test is the closeness of agreement between the scores and some other objective measure of that which the test is used to measure. The other measure is called the *criterion*. The coefficient of validity of a test is the coefficient of correlation between test scores and criterion scores. (Bingham, 1937: 214, italics in original)

Thus, within a number of prominent texts, validity came to be defined in terms of *observed* agreement between test scores and scores from a criterion measure, rather than in terms of *hypothetical* agreement between test scores and true proficiency. Understood in this manner, validity was reduced to no more nor less than an empirical correlation: the *coefficient of validity*. This was problematic as a definition,