Tracing the evolution of validity in educational measurement: past issues and contemporary challenges

Stuart Shaw CIE Research **and Victoria Crisp** Research Division

Introduction

Validity is not a simple concept in the context of educational measurement. Measuring the traits or attributes that a student has learnt during a course is not like measuring an objective property such as length or weight; measuring educational achievement is less direct. Yet, educational outcomes can have high stakes in terms of consequences (e.g. affecting access to further education), thus the validity of assessments is highly important.

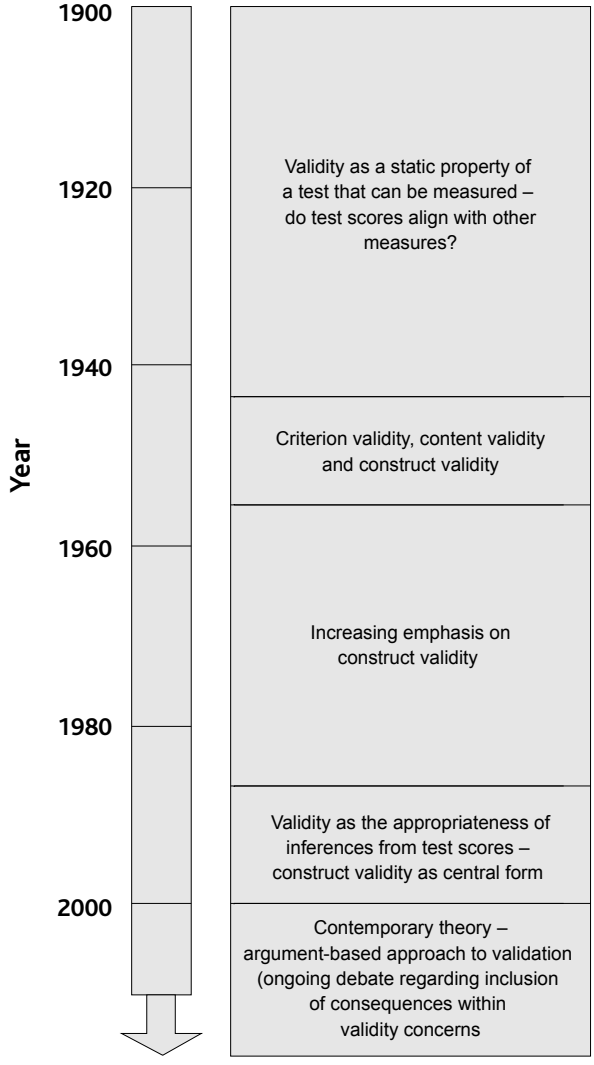

Figure 1 : Timeline of the evolution of validity theory

Tracing this trajectory of evolution, particularly through key documents such as the validity/validation chapter in editions of Educational Measurement (Cureton, 1951; Cronbach, 1971; Messick, 1989; Kane, 2006) and the Standards of Educational and Psychological Testing (AERA, APA and NCME, 1954/1955, 1966, 1974, 1985, 1999) has been important to us as part of work to develop an approach to validation for general assessments.

The concept of validity is not a new one. Conceptualisations of validity are apparent in the literature from around the turn of the twentieth century, and since that time, they have evolved significantly. Earliest perceptions of validity were that of a static property captured by a single statistic, usually an index of the correlation of test scores with some criterion (Binet, 1905; Pearson, 1896; Binet and Henri, 1899; Spearman, 1904). Through various re-conceptualisations, contemporary validity theory generally sees validity as about the appropriateness of the inferences and uses made from assessment outcomes, including some consideration of the consequences of test score use. This article traces the progress and changes in the theorisation of validity over time and the issues that led to these changes. A timeline of the evolution of validity is provided by Figure 1.

1900–1950: Early validity theory

Most early validity theory was located within a realist philosophy of science[1] and in terms of educational measurement couched within the highly scientific discourse of psychological testing, grounded as it was in a positivistic epistemology. During this time validity was conceived of as a statistical index, validity being evaluated in terms of how well the test scores predicted (or estimated) the criterion scores. The criterion measure was the value (or amount) of the attribute of interest. The attribute was assumed to have a definite value for each person and the objective of assessment was to arrive at an accurate estimation of the amount of attribute manifested. Thus validity was defined in terms of the accuracy of the estimate and validation was seen to require some criterion measure which was assumed to provide the 'real' value of the attribute of interest.

Early definitions placed emphasis on the test itself. Bingham defined validity from an operational perspective as the correlation of scores on a test with "some other objective measure of that which the test is used to measure" (Bingham, 1937, p.214) – a view shared by a number of well known measurement theorists at the time (including Cureton, 1951; Gulliksen, 1950) and most notably expressed by Guilford (1946), who said that "in a very general sense, a test is valid for anything with which it correlates" (p.429).

By the 1920s, tests were described as being valid for any criterion for which they provided accurate estimates (Thorndike, 1918; Bingham, 1937). For example, Kelley noted "the problem of validity is that of whether a test really measures what it purports to measure" (1927, p.14).This view prevailed throughout the first half of the twentieth century.

1. Scientific realism was developed largely as a reaction to logical positivism. Scientific realists claim that science aims at truth and that scientific theories should be regarded as true (or at least approximately true, or likely to be true).

1950s: Criterion-based, content-based and construct-based models of validity

During the 1950s, the concept of validity was refined to include the ability of a test to predict future performance with respect to external criteria (*criterion*), content area (*content*), or a theoretical construct (*construct*). In other words, validity was conceptualised as triune in nature comprising criterion, content and construct facets. Throughout this time, and even beyond, the tripartite division had become so widely embraced that Guion, writing in the 1980s, criticised how many took this structure 'on faith' and without questioning. He referred to it as "something of a holy trinity representing three different roads to psychometric salvation" (1980, p.386).

Criterion validity

The 1950s began with Cureton's sophisticated summary of conceptions of validity which he articulated prior to the advent of construct validity. Cureton (1951) stated that "The essential question of test validity is how well a test does the job it is employed to do" (p.621). Validity, he argued, "indicates how well the test serves the purpose for which it is use[d]" and, therefore, can be "defined in terms of correlation between the actual test scores and the 'true' criterion scores" (p.623). Essentially, he was arguing for the criterion model as offering the best solution to evaluating validity. This view was predicated on earlier conceptualisations of validity as a static property that could be measured in relation to a true criterion.

The criterion-based model, in which validity of the criterion and test scores were to be validated against the criterion scores, was helpful in a variety of applied scenarios, assuming that some suitable 'criterion' measure was available. Apart from being an objective measure, criterion-related evidence seemed relevant to the plausibility of proposed test score interpretations and uses.

In the 1954/1955 *Standards* (AERA, APA and NCME, 1954/1955) criterion validity was deconstructed into two forms of validity: concurrent validity and predictive validity. Concurrent validity made use of indirect measures which permitted validity estimates to be obtained concurrently with test scores, whilst predictive validity depended on a criterion of subsequent performance which could not be achieved concurrently with test scores. The 1966 *Standards* (AERA, APA and NCME, 1966) characterised criterion validity in the following way: criterion validity compared test scores with "one or more external variables considered to provide a direct measure of the characteristic or behaviour in question" (p.12).

However, there were issues with the criterion-based model which demanded a well-articulated and demonstrably valid criterion measure. Presupposing a criterion measure was available, questions about the validity of the criterion emerged. Unfortunately, the model was unable to provide a sound footing for validating the criterion. One possible solution was to employ a criterion measure involving some desired performance and then to interpret the scores in relation to that performance such that validity of the criterion could be accepted.

Content validity

Content validity methods focus on item content and the degree to which the test samples the 'universe' of relevant content. According to the 1966 *Standards* (AERA, APA and NCME, 1966), content validity demonstrated how well a test "samples the class of situations or subject matter about which conclusions are to be drawn." Much later, Messick (1989) described content-validity evidence as providing support for "the domain relevance and representativeness of the test instrument" (1989, p.17). It was deemed legitimate to extrapolate from an observed performance on a sample of assessment tasks from a domain as an estimation of generalised performance in the domain providing it could be demonstrated that the observed performances were representative of all assessment tasks and that the size of the sample was sufficiently large to control for sampling error (Guion, 1977).

However, content-validity evidence tended to be both subjective and confirmatory (based on judgement by experts who sometimes had a vested interest in the assessment) and did not involve test scores or performances on which scores were based. Consequently, it was difficult to justify conclusions about interpretation of test scores. Additionally, the content-based validity model proved to be problematic when used as grounds for arguing the validity of claims about cognitive processes or underlying theoretical constructs as the following quotes illustrate:

- "Judgments about content validity should be restricted to the operational, externally observable side of testing. Judgments about the subjects' internal processes state hypotheses, and these require empirical construct validation." (Cronbach, 1971, p.452)
- Content-based validity evidence provides "the domain relevance and representativeness of the test instrument" (Messick, 1989, p.17) but does not provide direct evidence for the "inferences to be made from test scores" (p.17).

It was becoming increasingly more necessary, given the shortcomings of both the criterion-based and content-based models, to develop a more sophisticated conceptualisation of validity.

Construct validity

Meehl and Challman (APA, 1954) first introduced the concept and terminology of construct validity, however, the concept was developed further by Cronbach and Meehl's (1955) seminal paper – 'Construct validity in psychological tests' – published in *Psychological Bulletin*. Much of their thinking had its origins in the hypothetico-deductive (HD) model of scientific theories (Suppe, 1977). Cronbach and Meehl began with the notion of a construct as "some postulated attribute of people assumed to be reflected in test performance" (1955, p.283), and asked the question whether the test was an adequate measure of the construct. According to Cronbach and Meehl, "determining what psychological constructs account for test performance is desirable for almost any test" (1955, p.282). They suggested that construct validity was an all-pervasive concern though they did not offer it as a general organising framework for validity. Cronbach and Meehl (1955) attempted to link theory and observation – a central tenet of construct validity, by constructing a nomological network. They proposed that the constructs that a test is intended to measure could be represented by a nomological network which included a theoretical framework (for what was being measured) and an empirical framework (for how it was going to be measured). Any associations between the two networks would need to be specified.

Thus, construct validity became the third 'type' of validity in thinking around this time. Construct validity served the purpose of inferring "the degree to which the individual possesses some hypothetical trait or quality (construct) ... that cannot be observed directly" by determining "the degree to which certain explanatory concepts or constructs account

for performance on the test ... through studies that check on the theory underlying the test" (AERA, APA and NCME, 1966, pp.12–13). The 1966 *Standards* distinguished construct validity from other forms of validity in the following way: "Construct validity is ordinarily studied when the tester wishes to increase his understanding of the psychological qualities being measured by the test ... Construct validity is relevant when the tester accepts no existing measure as a definitive criterion" (AERA, APA and NCME, 1966, p.13).

Essentially, construct validity attempted to make a link between assessment performance and pre-conceived theoretical explanations, in other words, to determine the consistency between observed performance on an assessment and its related underlying construct theory. One development of interest at this time came from Campbell and Fiske (1959) who proposed the multi-trait multi-method approach to validation. This included the introduction of two new concepts – convergent validity (the degree to which the test correlates with established tests or assessments purporting to measure similar constructs) and discriminant validity (the degree to which the test does not correlate with measures of different constructs). In practical terms this led to further validation methods involving the use of correlations between different measures, in order to evaluate the likelihood of similar constructs being assessed.

Important features of Cronbach and Meehl's (1955) construct model served as a general methodology for subsequent validation. They emphasised the need for extensive validation efforts, the need for an explicit statement of the proposed interpretation prior to evaluation and the need to challenge proposed interpretations and consider alternate interpretations.

Meehl and Challman (APA, 1954) and Cronbach and Meehl (1955) argued that construct validity offered an alternative to the criterion-based and content-based models. Shortly after the publication of Cronbach and Meehl's (1955) paper, Loevinger (1957) suggested that construct validity was an overriding concern subsuming the content and criterion models. She contended that only construct validity provided a scientifically useful basis for establishing validity. Her assertions foreshadowed Messick's unified view of validity by thirty years reflecting as it did the scientific principles of construct validity.

Kane (2006) asserts that construct validity is deeply based in logical positivistic assumptions which require a coherent and well-articulated theory from which to ground validity claims.

1955–1989: Evolution of the construct validity model

The model of construct validity posited by Cronbach and Meehl appeared to pave the way for validity thinking for the next decade or so, though the model was subject to significant refinement. In 1971, Cronbach wrote the second chapter on validity for *Educational Measurement* thereby adding to and developing Cureton's (1951) position. In his chapter, Cronbach gave construct validity more centrality in relation to the general conception of validity than had the 1966 Standards (AERA, APA and NCME, 1966). Whilst, he continued to maintain the relevance of the triune nature of validity, he likened validity research to the evaluation of a scientific theory as characterised in 'construct validity'. Cronbach argued that most educational assessments involved constructs: "whenever one classifies situations, persons, or responses, he uses constructs" (1971, p.462) and that, "Every time an educator asks 'but what does the instrument really measure?' he is calling for information on construct validity" (1971, p.463).

Cronbach defined validity in terms of interpretations and a range of potential uses and, like his predecessor Cureton, emphasised that validity is not an inherent property of a test but must be evaluated for each testing application:

> *Narrowly considered, validation is the process of examining the accuracy of a specific prediction or inference made from a test score ... More broadly, validation examines the soundness of all interpretations of a test – descriptive and explanatory interpretations as well as situation-bound predictions*. (Cronbach, 1971, p.443)

Within the compass of validity studies Cronbach also included evaluation of decisions and actions based on test scores as well as descriptive interpretations. Cronbach articulated a broad view of validation as involving the evaluation of the interpretations of assessment outcomes and argued that validation focuses on the "accuracy of a specific prediction or inference made from a test score" (1971, p.443). Cronbach (1971) also distinguished a number of approaches to validation, elaborating types of validation needed to support decision-oriented test use. He differentiated validity for selection from validity for placement and emphasised the need to integrate different kinds of validity evidence in evaluating the proposed interpretations and uses of test scores.

Echoing the sentiments expressed in the 1966 *Standards*, the 1974 *Standards* listed four types of validity associated with "four independent kinds of inferential interpretation" (1974, p.26) – predictive and concurrent validities, content validity and construct validity. At this time, the *Standards* explicitly stated validity in terms of its specific intended purpose and context: "No test is valid for all purposes or in all situations or for all groups of individuals" (APA, 1974, p.31).

Unlike criterion-based validation (in which the generation of a correlational index could support validity), or content-based validation, (in which experts attest to the validity of a test's content), construct validation necessitated extensive research effort. Methods employed in construct validation helped determine the link between observed assessment performance and its related construct theory – construct validation being associated with theoretical variables for which "there is no uniquely pertinent criterion to predict, nor is there a domain of content to sample" (Cronbach, 1971, p.462).

Educational measurement theorists throughout this period were beginning to understand that the test itself was not validated; rather, the focus of validation should be the inferences and decisions derived from scores on the test. Alongside this increased awareness was a recognition that multiple measures and multiple evidential sources should be taken into consideration when validating assessment inferences, especially in relation to complex domains.

Towards the end of the 1970s, there existed a tension between major validity theorists who regarded construct validity as dominant model pushing towards a more unified approach to the theory of validity (Cronbach, 1989; Guion; 1977, 1980; Messick, 1975, 1981; Tenopyr, 1977) and those (predominantly assessment users who saw the practical uses of predictive, content, and criterion validity) who continued to work from multiple validity frameworks.

Between the early 1950s and the late 1970s a practice had emerged whereby a 'toolkit' of different models was developed for validating

educational and psychological tests – different models to be employed for different assessments.

The 1980s

By the 1980s the construct model had been adopted as a general approach to validity (Anastasi, 1986; Embretson, 1983; Messick, 1980, 1988, 1989). Messick adopted a broadly defined version of the construct model as a unifying framework for validity. Messick perceived validity as a unified concept and that validity measures are not singular; rather, validity is an ongoing activity that relies on multiple evidence sources. According to Messick (1988, p.35): "from the perspective of validity as a unified concept, all educational and psychological measurement should be construct-referenced because construct interpretation undergirds all score-based inferences – not just those related to interpretive meaningfulness but also the content- and criterion-related inferences specific to applied decisions and actions based on test scores."

In his seminal treatise on validity in the third edition of *Educational Measurement*, Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (1989, p.13). This definition resonates with the definition provided in the most recent version of the *Standards* (AERA, APA and NCME, 1999). Messick (1989) conceptualised validity in terms of value implications and social consequences of testing outcomes. He emphasised validity as an evaluative process focusing on inferences derived from assessment scores (not the assessment itself) and the actions resulting from those inferences. Messick argued that validity extends beyond test score meaning and includes aspects related to score relevance and utility, value implications, and social consequences.[2]

In challenging the 'unholy trinity' of validity, Messick perceived score meaning and construct validity as the underlying objective of all test validation: "validation is a matter of making the most reasonable case to guide both current use of the test and current research to advance understanding of what the test scores mean … To validate an interpretive inference is to ascertain the degree to which multiple lines of evidence are consonant with the inference, while establishing that alternative inferences are less well supported" (1989, p.13).

Messick argued that validation "embraces all of the experimental, statistical, and philosophical means by which hypotheses and scientific theories are evaluated" (1989, p.14) and entails:

- determining "the degree to which multiple lines of evidence are consonant with the inference, while establishing that alternative inferences are less well supported" (1989, p.13).
- "appraisals of the relevance and utility of test scores for particular applied purposes and of the social consequences of using the scores for applied decision making" (1989, p.13).

It is important to stress that validity conceptualised as a unified view did not in any way diminish content or criterion sources of evidence but instead subsumed them in an attempt to build a robust argument for validity. Moreover, the unified approach permitted a fusion of competing theories and validation methodologies. A key point was the idea that a unified, though multi-faceted concept of validity, constituted the foundation for contemporary validity theory.

Contemporary validity theory

Ratcliffe (1983) observed that "quite different notions of what constitutes validity have enjoyed the status of dominant paradigm at different times, in different historical contexts, and under different prevailing modes of thought and epistemology" (p.158). Echoing Ratcliffe's sentiments, Moss, Girard and Haniford (2006) suggest that validity theory can be understood "as an intellectual framework or set of conceptual tools that shapes both our understanding and our actions" (p.109) and as "the representation of an epistemology – a philosophical stance on the nature and justification of knowledge claims – which entails a philosophy of science" (p.110). The epistemological shift in validity theory from a positivistic[3] to a post-positivistic orientation[4] (Moss *et al*., 2006) – described elsewhere by Geisinger (1992) as a 'metamorphosis' – has brought about a variety of epistemological and methodological perspectives within contemporary validity theory (DeLuca, 2009).

In the fourth and latest edition of *Educational Measurement*, Kane (2006) calls for multi-perspective validity arguments to justify test use. Citing House's (1980) logic of evaluation and Cronbach's (1988) earlier work on validation as an evaluation argument, Kane proposes an argument-based approach to validity. Kane's validation framework is congruent with the approach to validation suggested by the current version of *Standards* (AERA, APA and NCME, 1999) and resonates with Messick's 1989 chapter on validity.

Kane (2006) perceives the validation process as the assembly of an extensive argument (or justification) for the claims that are made about an assessment. According to Kane, "to validate a proposed interpretation or use of test scores is to evaluate the rationale for this interpretation or use. The evidence needed for validation necessarily depends on the claims being made. Therefore, validation requires a clear statement of the proposed interpretations and uses" (2006, p.23). Cronbach also conceptualised validity arguments as serving an evaluative function stating that, "validation of test or test use is evaluation" (1988, p.4).

Kane proposed that any validation activity should necessarily entail both an *interpretive* argument (in which a network of inferences and assumptions which lead from scores to decisions is explicated) and a *validity* argument (in which adequate support for each of the inferences and assumptions in the interpretive argument is provided and plausible alternative interpretations are considered).

An argument-based approach to validation is perceived to constitute a compromise between complicated validity theory and a requirement to present a case for the defensibility of using a test for a specified purpose. The force of an argument-based approach to validation is that it:

- ensures that the task of validating inferences is both scientifically sound and logistically manageable;
- provides guidance in apportioning research resource;
- enables estimates of progress in the validation effort to be made;
- and facilitates identification of the various sources of validity evidence that would support or refute the inferences specified on the basis of test scores.

2. A criticism of construct validity as the framework for a unified model of validation was that it did not provide clear guidance for the validation of a test score interpretation or use.
3. The positivistic position assumes a reality that is independent of human perception and therefore draws a distinction between facts and values (Denzin & Lincoln, 2008).
4. The post-positivistic mode of inquiry recognises truth as socially constructed, situational and subjective (Denzin & Lincoln, 2008).

To claim that an interpretation or use of a test is valid is "to claim that the interpretative argument is coherent, that its inferences are reasonable, and that its assumptions are plausible" (Kane, 2006, p.23). In terms of Kane's framework, validation activity requires sufficient evidence that: the test actually measures what it claims to measure; the test scores demonstrate reliability; and that the test scores manifest associations with other variables in a way that is compatible with its predicted properties.

The role of consequences in validity

The most recent version of the *Standards* (AERA, APA and NCME, 1999) identifies five sources of validity evidence, one of which is "evidence based on consequences of testing" (1999, p.16).[5] Describing consequences, the *Standards* "distinguish between evidence that is directly relevant to validity and evidence that may inform decisions about social policy that falls outside the realm of validity" (1999, p.16). That the role of consequences should be included in the *Standards* as a potential source of validity evidence is undoubtedly a result of Messick's (1989) hugely influential chapter in which he formalises the consequential bases of test interpretation and test use. Messick's (1989) integration of both evidential and consequential sources of evidence have served to appreciably widen the compass of validity inquiry by including social and value-laden aspects of assessments thereby extending traditional measurement boundaries into issues relating to policy – what Kane (2001) has termed the prescriptive part of a validity argument. This has necessitated the requirement for evidence about the social consequences of test use (Cronbach, 1988; Messick, 1989, 1994; Shepard, 1993; Linn, 1997). However, whether Messick's definition of validity included evidence about all consequences of assessment as validity is fiercely contested. Even Shepard, an advocate of consequential validity, acknowledges that ''there is a great deal more in what Cronbach and Messick have suggested than is acknowledged or accepted by the field'' (1993, p. 406).

The role of consequences in testing has become a highly controversial issue within contemporary validity debate (Crocker, 1997; Brennan, 2006). Brennan states, "since it is now almost universally agreed that validity has to do with the proposed interpretations and uses of test scores, it necessarily follows that consequences are a part of validity" (2006, p. 8). However, there is considerable disagreement regarding the role that the consequences of test score use plays in validity theory. The importance of the debate is most clearly illustrated by the fact that two entire issues of the journal *Educational Measurement: Issues and Practice* were given over to such concerns in 1997 and 1998.

Of course the role of consequences in testing is not new. Cureton (1951) acknowledged consequences as being a part of validity in his chapter and Kane (2006) maintains that consequences have always played an integral role in validation. There exists within the educational measurement community, therefore, general agreement that evaluating consequences is important. What is contentious, however, is the validation of both intended consequences (claimed outcomes) and unintended or negative consequences of test use. Since consequences reflect the effects or impacts of test usage, evaluating intended consequences is ostensibly an attempt to evaluate the extent to which a test fulfills its specified purpose or proposed use. For a full evaluative treatment of all consequences to be complete, analysis of evidence would require monumental validation effort especially if it is to include an exploration of unintended consequences.

Some measurement theorists (Maguire, Hattie and Haig, 1994; Crocker, 1997; Green, 1998; Mehrens, 1997; Popham, 1997; Borsboom, Mellenbergh and van Heerden, 2004) have argued for a limited and more technical definition of validity that emphasises the descriptive interpretation of scores. Whilst they suggest that consequences are crucial to social research they nevertheless categorise them as being outside validity theory. According to Maguire *et al*., "Consequences should be moved out from the umbrella of construct validity and into the arena of informed social debate and formulated into ethical guidelines" (1994, p.115). Others, however, embrace a broader view of validity arguing that assessments should be contextualised within their consequential outcomes (e.g. Linn, 1997; Messick, 1989; Moss, 1998; Shepard, 1997; Kane, 2001).

Summary

Within the sphere of educational assessment there is now broad agreement regarding Messick's (1989) definition of validity as about the appropriateness of the inferences and uses of assessment outcomes (though this is by no means universal, see for example, Borsboom, 2006; Borsboom, Mellenbergh and van Heerden, 2004; Lissitz and Samuelsen, 2007). Validition is perceived by Kane (2006) to be a judgement of the degree to which arguments support those proposed interpretations and uses. Following an extensive review of the literature, Sireci (2007, 2009) summarises the fundamental features of validity in the following ways (2007, p.477):

- validity is not an inherent property of a test but refers to the specified uses of a test for a particular purpose;
- validity refers to the proposed interpretations or actions that are made on the basis of test scores;
- in order to evaluate both the usefulness and appropriateness of a test for a particular purpose multiple sources of evidence are required;
- sufficient evidence must be collected to defend the use of the test for a particular intended purpose;
- the evaluation of validity is neither static nor a one-time event but a continuing process.

Messick (1989) argued that "validity is an evolving property and validation is a continuing process" (p.13). The contemporary conceptualisation of validity cannot be considered definitive, but as the current most accepted notion. This, and particularly the role of consequences as part of validity, is likely to continue to evolve over time.

5. The other sources of validity evidence include *test content*; *response processes*; *internal structure*; and *relations to other variables*

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1954/1955). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques [supplement]. *Psychological Bulletin*, **51**, 2, Pt.2, 201–238.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, **37**, 1–15.

Binet, A. (1905). New methods for the diagnosis of the intellectual level of subnormals. *L'Année Psychologique*, **12**, 191–244.

Binet, A., & Henri, B. (1899). La psychologic individuelle. *Amiee Psychol*., **2**, 411–465.

Bingham, W. V. (1937). *Aptitudes and aptitude testing*. New York: Harper.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, **71**, 425–440.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, **111**, 1061–1071.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In: R. L. Brennan (Ed.), *Educational Measurement*, (4th ed.), 1–16. Westport, CT: Praeger.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, **56**, 81–105.

Crocker, L. (1997). The great validity debate. *Educational Measurement: Issues and Practice*, **16**, 2, 4.

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

Cronbach, L. J. (1988). Five perspectives on validity argument. In: H. Wainer (Ed.), *Test validity*, 3–17. Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In: R. L. Linn (Ed.), *Intelligence: Measurement, theory and public policy*, 147–171. Urbana: University of Illinois Press.

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

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

DeLuca, C. (2009). *Contemporary Validity Theory in Educational Assessment: Integrating an Interpretivistic Approach through Case Study Methodology*. Paper presented at the annual meeting of the Canadian Society for the Study of Education, Ottawa, Ontario, May 2009.

Denzin, N. K., & Lincoln, Y. S. (2008). *Strategies of Qualitative Inquiry*. (3rd ed.). Thousand Oaks, CA: Sage.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, **93**, 179–197.

Geisinger, K. F. (Ed.) (1992). *The psychological testing of Hispanics*. Washington, DC: APA.

Goodwin, L. D. (2002). Changing conceptions of measurement validity: An update on the new standards. *Journal of Nursing Education*, **41**, 100–106.

Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement: Issues and Practice*, **17**, 2, 16–19, 34.

Guion, R. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement*, **1**, 1–10.

Guion, R. M. (1980). On trinitarian conceptions of validity. *Professional Psychology*, **11**, 385–398.

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

Gulliksen, H. (1950). Intrinsic validity. *American Psychologist*, **5**, 511–517.

House, E. T. (1980). *Evaluating with validity*. Beverly Hills, CA: Sage Publications.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, **38**, 4, 319–342.

Kane, M. T. (2006). Validation. In: R. L. Brennan (Ed.), *Educational Measurement*. (4th ed.). 17–64. Westport, CT: American Council on Education/Praeger.

Kelley, T. L. (1927). *Interpretation of Educational Measurements*. New York: New World Book Company.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, **16**, 2, 14–16.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, **36**, 437–448.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, **3** (Monograph Supplement 9), 635–694.

Maguire, T., Hattie, J., & Haig, B. (1994). Construct validity and achievement assessment. *Alberta Journal of Educational Research*, **40**, 109–126.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, **16**, 2, 16–18.

Messick, S. (1975). The standard program: Meaning and values in measurement and evaluation. *American Psychologist*, **30**, 955–966.

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

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher*, **10**, 9–20.

Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In: H. Wainer and H. Braun (Eds.), *Test validity*, 33–45. Hillsdale, NJ: Lawrence Erlbaum.

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher,* **23**, 2, 13–23.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, **17**, 2, 6–12.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education*, **30**, 109–162.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society A*, **187**, 253–318.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice*, **16**, 2, 9–13.

Ratcliffe, J. W. (1983). Notions of validity in qualitative research methodology. *Knowledge: Creation, Diffusion, Utilization*, **5**, 147–167.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, **19**, 405–450.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, **16**, 2, 5–8, 13, 24.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, **36**, 477–481.

Sireci, S. G. (2009). Packing and Unpacking Sources of Validity Evidence. In: Lissitz, R. W. (Ed.), *The Concept of Validity: Revisions, New Directions, and Applications*. IAP: Charlotte, NC.

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

Suppe, P. (1977). *The structure of scientific theories*. Urbana, IL: University of Illinois Press.

Tenopyr, M. L. (1977). Content-construct confusion. *Personnel Psychology*, **30**, 47–54.

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurements of educational products. *National Society for the Study of Educational Products: Seventeenth Yearbook*. 16–24.