EL.LE ISSN 2280-6792
Vol. 1 — Num. 1 — Marzo 2012

New views of validity in language testing

Claudia D'Este

Abstract Language testing has been defined as one of the core areas of applied linguistics because it tackles two of its fundamental issues: the need to define and reflect on the appropriateness of Second Language Acquisition models and constructs by data analysis from language tests and the importance of facing the ethical challenge deriving from the social and political role language tests play nowadays. Language testing has thus a twofold impact in a variety of contexts. In the first instance, it constitutes a scientific impulse for which research is needed to develop and implement the technical design of tests. Secondly, language testing has also become subject of debating because the use and interpretation of test results introduce ethical issues concerning the concept of 'fairness' in the construction, administration, evaluation and interpretation of language tests. In fact, language tests are always designed and used to make decisions on the basis of a process in which information about test takers is gathered from an observed performance under test conditions. This inevitably leads to the development of codes of ethics in educational testing environments and to the elaboration of theories of validity and validation.

1 Introduction

Language testing has been defined as one of the core areas of applied linguistics because it tackles two of its fundamental issues: the need to define and reflect on the appropriateness of Second Language Acquisition models and constructs through data analysis from language tests, and the importance of facing the ethical challenge deriving from the social and political role language tests play nowadays.

Language testing has thus a twofold impact in a variety of contexts. In the first instance, it constitutes a scientific impulse for which research is needed to provide accurate measures of precise abilities. Secondly, language testing has also become a subject of debate because the use and interpretation of test results introduces ethical issues concerning the concept of 'fairness' in the construction, administration, evaluation and interpretation of language tests: the powerful effect of the '(mis)use' of the test that might have «harmful unintended or intended consequences» for test takers or society (Fulcher 1999).

In fact, language tests are always designed and used to make decisions on the basis of a process in which information about test takers is gathered from an observed performance under test conditions. This inevitably leads to the development of codes of ethics in educational testing environments and to the elaboration of theories of validity and validation in general education and in language testing.

Starting from the conceptualization of different types of validity borrowed from psychometrics and transposed into language testing theory by Lado (1961), the real turning point was marked in 1985 with the issue of *The Standards for Educational and Psychological Testing* developed jointly by the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The Standards introduced a framework of reference intended as a comprehensive basis for evaluating educational tests. They were strongly influenced by the theories on validity of Messick (1989), who introduced a unified concept of validity in which 'consequential validity' is the facet concerned with the social consequences of test use and how test interpretations are arrived at. Another related area of research on validity which has recently attracted a great deal of attention is that of 'washback'. Research on washback investigates the relationship between testing and teaching, between test use and the society in which it is used.

In the light of the above considerations, this paper examines new issues on the concepts of validity and validation procedures in language testing. The arguments I shall put forward are relevant to new approaches and their potential connections to operational testing situations by analysing different validation methods. For the purpose of this paper, I will first scrutinise different definitions of validity and give an overview of the educational and psychological origins of theories of validity and validation. I will then introduce Messick's views on validity and Bachman's model of test validity. Lastly, I will address issues on validation methods and scope with a focus on problems arising from the need to investigate aspects relevant to the test users and context.

2 Definitions of test validity

The meaning of test validity has undergone a metamorphosis over the last fifty years. Early definitions of validity put the utmost emphasis on the test itself as validity was considered a static property. A test was considered to be either valid or not as evidenced by the correlations between the test and some other «external» criterion measure (Goodwin, Leech 2003). The concept of validity applied to testing was first investigated by psychometrics (the field of study concerned with the theory and technique of educational and psychological measurement: validity is the degree to which a test measures what it is designed to measure.

In 1955, Cronbach and Meehl identified four types of validity: predictive validity, concurrent validity, content validity, and construct validity. The first two types of validity are considered together as 'criterion-oriented validation' procedures because they deal with some criterion that the investigator is primarily interested in predicting when s/he administers the test. **Predictive validity** is studied when the criterion is obtained some time after the test has been administered. **Concurrent validity** is examined when test score and criterion score are determined at essentially the same time and can be studied when one test is proposed as a substitute for another. **Content validity** represents the extent to which the items of a test are appropriate to the content domain of the test, and it is established by showing that the test items are a good example of a universe in which the investigator is interested. **Construct validity** is involved whenever we want to interpret a test as a measure of some attribute or quality which is not 'operationally' defined. 'Construct' is the core validity and it is studied when the tester needs to demonstrate that an element is valid by relating it to another element that is supposed to be valid and when the construct underlying the test is more important than either the test behaviour or the scores on the criteria.

In 1961, Robert Lado provided the first significant contribution to language testing by applying the term 'validity' to language testing. He conducted his research on the basis of a question-statement which summarizes his concept of validity: «Does a test measure what it is supposed to measure? If it does, it is valid». Validity is thus considered as one of the most important qualities of a language test together with **reliability**. Validity is «a matter of relevance»: a test is considered valid when test content and test conditions are relevant and there are no «irrelevant problems which are more difficult than the problems being tested» (Lado 1961, p. 321). Summarizing the work by Lado (1961) and Davies (1968), validity can be established in several ways: face validity (to decide whether the test is valid by simple inspection), validity by content (to check the validity of content of the items), control of extraneous factors (typical of foreign language testing when non native speakers introduce elements of their language and culture in the test-taking process), validation of the conditions required to answer the test items and empirical validation (to compare the scores on the test with some other criterion whose validity is self-evident). Lado also pointed to the importance of reporting the validity of a test in expectancy tables and as a correlation coefficient.

Campbell and Fiske (1959) introduced discrete types of validity and the need for different types of validity evidence: a **multimethod-multitrait approach** to validation, which included the introduction of **convergent** and **discriminant** types of validity. **Convergent validity** demonstrates that measures that should be related are in reality related whilst **discri-**

minant validity shows that measures that should **not** be related are in reality **not** related.

The concepts of 'internal' and 'external validity' are also used by Campbell and Stanley (1966) in the field of experimental design to describe and investigate the causes of the results of a particular study and the way in which independent and dependent variables are linked together in a cause-effect relationship. **Internal validity** is the *sine qua non*, essential validity and it is specific to the experiment, whilst **external validity** asks the questions of generalizability, or to what extent the findings of an experiment can be applied to different groups or settings and times.

The landmark publication of the 1966 Standards (APA, AERA, NCME 1966) shifted the focus onto the meaning of validity to use, and validity was thus defined as the extent to which a test produced information that was useful for a specific purpose. Three categories (strongly linked to the 'trinitarian' view of validity first presented by Cronbach and Meehl in 1955) emerged: content validity, criterion-related validity (which included concurrent and predictive validities), and construct validity.

Alderson, Clapham and Wall (1995) introduced the terms 'internal and external validity' in language testing. The first type of validity refers to studies of the perceived content of the test and its perceived effect, whilst the second type relates to studies comparing students test scores with measures of their ability collected from outside the test (Alderson, Clapham, Wall 1995, p. 171). Moreover, they introduce external validity into the concept of 'criterion validity' as defined by the American Psychology Association in 1985.

In fact, it was during the 1980s that another fundamental contribution in the conceptualization of validity was made. The 1985 Standards (AERA, APA, NCME 1985) described validity as «the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores», and test validation as «the process of accumulating evidence to support such inferences» (AERA, APA, NCME 1985, p. 9). The Standards are based on the premises that effective testing and assessment require test developers and users to be 'knowledgeable' about validity, reliability and other measurement issues, and thus refer to validity as to «the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests» (AERA, APA, NCME 1985, p. 9).

Shortly after the issue of the 1985 Standards, Messick (1989) introduced a new **unified** validity framework based on two main **interconnected** facets in order to highlight the inferences and decisions made from test scores. One facet is «the source of justification of the testing, being based on appraisal of either evidence or consequence», whilst the other facet is represented «by function or the outcome of the testing, being either interpretation or use» (Messick, 1989, p. 20). Messick's view was revolutionary because it contrasted with the traditional definitions of validity.

	Test interpretation	**Test use**
Evidential basis	Construct validity	Construct validity + Relevance / Utility
Consequential basis	Construct validity + Value implications	Construct validation + Relevance / Utility + Value implications + Social consequences

Table 1 (from Messick 1989, p. 20).

The fundamental aspects of test validation were represented in the form of a matrix (Table 1) where 'test interpretation' involves gathering evidence and consequence of test validity outside a specific context in which the test is used, and 'test use' refers to the real use of the test in a well-defined context. The most important aspect of this distinction was the analysis of the potential misuse of a test, which may be well founded on a theory of the abilities it intends to measure, but which might be not appropriate in a particular context.

For Messick (1989, p. 245), validity is thus

> an overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other methods of assessment.

It is intended as a property of the test scores, and what is to be validated are the inferences deriving from test scores interpretation and the consequent actions or decisions that the interpretation involves: 'consequential validity'.

Messick devotes a superordinate role to 'construct validity' in his framework. In a more recent article, he stated that «test validation is empirical evaluation of the meaning and consequences of measurement, taking into account extraneous factors in the applied setting that might erode or promote the validity of local score interpretation and use» (Messick 1996, p. 246). The key concept is that of 'score meaning', which is defined as a «**construction** that makes theoretical sense out of both the performance regularities summarized by the score and its pattern of relationships with other variables, the psychometric literature views the fundamental issue as **construct** validity» (Messick 1996, p. 246). Messick developed the concept of 'consequential validity' as a unified notion of validity, suggesting that the social consequences of test use must be estimated by putting together all the elements introduced in Figure 1 to make a judgement about the long-term effects deriving from the use of a test. In this respect, he introduces the concept of 'washback' of tests in educational practices.

Detailed implications of Messick's views in language testing have been outlined by Bachman (1990), who claimed that the validity of a given use of test scores is the outcome of a complex process that must include «the analysis of the evidence supporting that interpretation or use, the ethical values which are the basis for the interpretation or use but also the test takers' performance» (Bachman 1990, p. 237). In order to investigate the concepts of validity and validation, reliability must also be taken into account as the complementary aspect of interpreting and distinguishing different reasons for variance in test scores. Reliability is the means by which we can investigate the variance due to factors involved in measurement or in test scores. In validating a test, we must consider other sources of variance, and must utilize a theory of ability to identify these sources (Bachman 1990, p. 239).

Starting from Messick's 'progressive matrix', Bachman focused on construct validity (the essential component of each cell of Messick's framework and an indicator of the individual's ability), and on the «value implications of interpreting the score in a particular way» by considering, for instance, the theories of language and the relevant educational and social ideologies we attach to the score interpretation (Bachman 1990, p. 243).

Bachman drew on Messick's theories and developed an extended framework of validity as illustrated in Table 2. He started from the analysis of the evidential 'basis of validity', which he refers to as the gathering of complementary types of evidence into the process of validation to support the relationship between test score and interpretation and use. The collection of evidence is necessary to show that a test is an adequate indicator of a given ability.

Three general types of evidence are to be collected in support of a particular test use: **content relevance and coverage** (the domain specification upon which a test is based), **criterion relatedness** (demonstrating a relationship between test scores and some criterion which is also an indicator of the ability tested), and **meaningfulness of construct** (concerning the extent to which performance on tests is consistent with predictions that we make on the basis of a theory of abilities, or constructs).

As far as the **consequential (or ethical) basis of validity** is concerned, Bachman argued that tests are not designed and used in a «value-free psychometric test-tube» but that they meet the needs of an educational system or of the whole society for which we must assume the potential consequences of testing. In considering the use of a test and the validity of the use of test scores, there is a shift from the scientific demonstration

Evidential basis of validity

Content relevance and content coverage (content validity)		Criterion relatedness (criterion validity)		Meaningfulness of construct (construct validity)	
Relevance	Coverage	Concurrent criterion relatedness (concurrent validity)	Predictive utility (predictive validity)	Logical analysis	Empirical investigation
- Ability domain - Test method facets	- Behavioural domain - Task analysis	- Differences in test performances - Correlations	- Use of information on criterion relatedness	- Theoretical and operational definition of constructs	- Observation of behavior (scores) - Correlational evid. - Experimental evid. - Analysis of the processes underlying test perfomance

Consequential (or ethical) basis of validity

Construct validity	Value system	Practical usefulness	Misuse of the test and consequences

Table 2. Evidential and consequential basis of validity (adapted from Bachman 1990).

of empirical and logical evidence to the arena of public policy (Bachman 1990, p. 281).

The category of 'consequential basis of validity' is divided into four areas that must be considered in the interpretation and use of test scores. Construct validity is still the central focus as it provides the evidence that

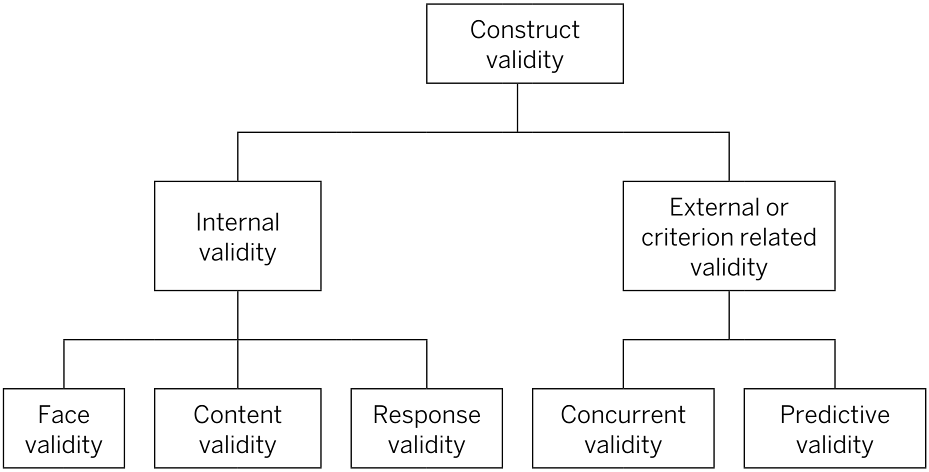

Figure 1. Categories of validity (adapted from Alderson, Clapham, Wall 1995).

supports the particular interpretation of the test we want to make. Another important area is that of 'the value system' that informs the particular test use. All the people involved in the testing process (developers, test takers and users) have their own value systems that may overlap in part or completely or be in opposition to each other in a given testing situation. Furthermore, it is important evidence to support the practical usefulness of a test and the analysis and collection of information that will determine the possible consequences of test use. It is the responsibility of test developers and test users

> to provide as complete evidence as possible that the tests that are used are valid indicators of the abilities of interest and that these abilities are appropriate to the intended use, and then to insist that this evidence be used in the determination of test use (Bachman 1990, p. 285).

Gipps (1994) considers that 'consequential validity' represents a shift from: «a purely technical perspective to a test-use perspective»: an ethical perspective (Gipps 1994, p. 146). She also refers to the evidence available to support test interpretation and potential consequences of test use, among which she includes the **washback** on teaching and the curriculum, which «are long-established consequences of assessment, particularly high-stakes testing» (Gipps 1994, p. 146).

Alderson, Clapham and Wall (1995) agree with Bachman: it is correct to define different methods of assessing validity because it is best to validate a test in as many ways as possible. They group validity into three main categories: **rational** or **content validity** which depends on a logical analysis to show that the content of the test is a good sample of the relevant language skill, **empirical validity** which is based on empirical and statistical evidence and **construct validity** which deals with what the scores mean. They also draw on the concept of **internal** and **external validity** to describe some methods of assessing validity which contribute to a superordinate form of validity: **construct validity** (Figure 1). **Internal validity** is represented by: **face validity** or how non-testers comment on the value of the test, **content validity** in which experts judge the test and **response validity** in order to collect information on how individuals respond to test items through introspective data. **External validity** is based on **concurrent validity** in which we compare test scores with other measures from the same test-taker taken at about the same time of the test and predictive validity when external measures are gathered some time after the test has been administered.

As we can see from the above overview, the concept of validity has evolved over the past decades, but it still appears powerful yet uncertain because it concerns the truth value of a test and its scores. It is also difficult to understand and analyse because of its abstract nature. According to Davies and Elder (2005), validity can be defined, established and measured only in an operational way, but when we want to put it in practice, we must turn our discussion into a consideration of what is the validation process.

3 The validation process: focus and methods

Test validation has been described as an exacting process that requires many types of evidence, analyses and interpretation (Cumming 1996). It aims at investigating the meaningfulness and defensibility of the inferences we make about individuals based on their test performance (McNamara 2004). The process of validation establishes the relationship between the claims of the test and the evidence in support of these claims. The scope is very important and many methods and frameworks are used to provide evidence for the assumption on which the inferences of a test are presented. Validation frameworks in language testing have been influenced by psychometric and statistical methods, by Second Language Acquisition theories and by Psychology. Nonetheless, the search for a reliable validation framework is an ongoing one.

Weir (2005) proposes a model of validation process in which test developers should work to generate evidence of the validity of a test from a different perspectives. His framework is «socio-cognitive in that the abilities to be tested are demonstrated by the mental processing of the candidate

[…]; equally, the use of language in performing tasks is viewed as a social rather than a purely linguistic phenomenon» (Weir 2005, p. 3).

Weir sees all the elements linked to each other through a symbiotic relationship: for example, 'context validity' (the traditional content validity), 'theory-based validity' and 'scoring validity' (an umbrella term encompassing various aspects of reliability) constitute 'construct validity'. Weir introduces the five key elements of his validation framework (context validity, theory-based validity, scoring validity, consequential validity, criterion-related validity) into socio-cognitive models for validating reading, listening, speaking and writing tests. He proposes different frameworks for each of the four skills. In all of them, test takers and their characteristics (physical/physiological, psychological and experiential) play a fundamental role because they are considered as elements relevant to test design. The test takers' characteristics are interrelated with 'context' and 'theory-based validity'. 'Scoring validity' parameters allow the evaluation of response and, finally, on the basis of 'consequential' and 'criterion related validity', the score/grade is established.

In a more recent article, Shaw and Weir (2007) provide a framework for conceptualizing writing test performance from which they derive fundamental questions that «anyone intending to take a particular test or to use scores from that test would be advised to ask of the test developers in order to be confident that the nature and quality of the test matches up to their requirements» (Shaw, Weir 2007, p. 5). These questions represent a comprehensive approach to a writing test's validation, and are the source of all the evidence to be collected on each of the components of this framework in order to improve the validity of the test.

In another important article on validation, Xi (2008) elaborates a graph, inspired by the theories of Cohen (1999), Kane and Crooks (1999) and Bachman (2005), which shows a network of inferences linking test performance to a score-based interpretation and use. She starts from Kane's theories (1999), for which validation is basically a two-stage process including the construction of an 'interpretative argument' and the development and evaluation of a 'validity argument'. The interpretative argument encompasses a logical analysis of the link between inferences from a test performance and relevant decisions in the light of the test's premises. If the network of inferences is supported by true assumptions, a sample of test performance and its corresponding score becomes more significant, and thus a scored-based decision has full justification.

The validity argument allows the evaluation of the interpretative argument using theoretical and empirical evidence. Xi also applies Bachman's (2005) adaptation of the validity argument which distinguishes a descriptive part (from test performance to interpretation) and a prescriptive part (from interpretation to decision).

Starting from the network of inferences, empirical methods of validation are illustrated. They are divided into groups according to the kind of support they provide for the inferential links: evaluation, generalization explanation, extrapolation and utilization (Xi 2008, p. 182).

Evaluation inferences are supported by evidence regarding the conditions of test administration and the attention paid to the development and application of the scoring rubrics. According to Xi (2008, p. 183), methods of validation may consist of:

1. Impact of test conditions on test performance: it is important to find out whether construct irrelevant factors influence test scores such as computer literacy in computer based test, differences between face-to-face or tape-mediated version of an oral test.
2. Scoring rubrics: rubrics play a fundamental role in a language test and if they do not mirror the relevant skills we may incur wrong scores. It is necessary to develop good rubrics by analysing samples of test discourse taken from rater verbal protocols or by validating rating scales.
3. Systematic rater bias studies: inconsistencies in assessment might be caused by subjective raters scoring. Methods for collecting evidence are: analysis of variance and multifaceted measurement to investigate «the systematic effect» of raters backgrounds on the scores, rater verbal protocols, questionnaires or interviews to investigate «rater orientations and decisions processes», rater self-reported data and the use of automated engines for scoring constructed response items.

In order to support **generalization** inferences, evidence can be gathered through:

1. Score reliability classical test theory (CTT), overall estimates of scores reliability by generalizability (G) theory and multifaceted Rash measurement. G theory informs us about the effects of the facets «such as raters or tasks and their dependability» while multifaceted Rasch measurement provides data on «the influence of individual raters, tasks and specific combinations of raters, tasks, and persons on the overall score reliability» (Xi 2008, p. 184).

   In test tasks, abilities and processes are engaged in real life language tasks justified by a domain theory which can account for performance in the domain. The **explanation** inferences (Xi 2008, pp. 184-187) are based on these assumptions, and different methods can be used to collect evidence about it.
2. Correlational or covariance structure analyses: they analyse the empirical relationship among items of a test or between the test and other measures of similar or different constructs to determine if these relationships are consistent with theoretical expectations (Xi 2008, pp. 184-187).

3. Experimental studies: instructions or learning interventions are planned and testing conditions and task characteristics are manipulated in a systematic way in order to emphasize the relationship between task performance and task difficulty, or to disambiguate a task feature suspected to be construct-irrelevant (Xi 2008, pp. 184-187).
4. Group difference studies: they take into account the possibility that groups with certain backgrounds and characteristics should differ with respect to the construct being measured (Xi 2008, pp. 184-187).
5. Self-report data on processes: verbal protocols and self-report data can be useful in finding out whether the test engages the abilities which it intends to assess, whether construct-relevant or construct-irrelevant task characteristics influence performance (Xi 2008, pp. 184-187).
6. Analysis of test language: discourse-based analyses of test language describe test-taking processes and strategies (Xi 2008, pp. 184-187).
7. Questionnaires and interviews: they are very useful tools to explore test-taking processes, strategies and reactions to test tasks and the whole test (Xi 2008, pp. 184-187).
8. Observational data on test-taking processes: together with post-test interviews, they reveal strategies and processes engaged by examinees, and possible bias introduced by the structure of a test (Xi 2008, pp. 184-187).
9. Logical analysis of test tasks: it combines judgemental analysis of the skills and the processes required by test tasks, interpretation of factors and of performance differences across groups or experimental conditions (Xi 2008, pp. 184-187).

Two kinds of evidence support the **extrapolation** inference: judgemental evidence (to demonstrate the domain representativeness of test tasks samples) and empirical evidence (to prove high correlation between test scores and scores on criterion measures).

Needs analysis and corpus-based studies are generally used because it is fundamental to specify the domain and logical analysis of the task content by content specialists, and to check the correspondence between the language used in test materials and real language use (Xi 2008, pp. 187-188).

Methods to gather evidence for the explanation and the extrapolation of inferences are fundamental in supporting the relevance of an assessment for its intended use. According to Xi (2008, pp. 188-189), the methods supporting **utilization** inferences are based on the examination of score reports and other materials provided to users, on the decision-making processes, and on the consequences of test use:

1. Score reporting practices and other materials provided to users: these represent the sole information on which test users (such as employers

or institutions) base their decisions so they must be useful and sufficient for decision-making (Xi 2008, pp. 188-189).

2. Decision-making processes: inappropriate cut score models or cut score requirements may turn into inappropriate decisions so it is important to set, verify and disclose the cut scores for minimal requirements (Xi 2008, pp. 188-189).
3. Consequences of using the assessment and making intended decisions: it mostly focused on washback, the impact of language tests on teaching and learning (Xi 2008, pp. 188-189).

4 Problems, operational implication and provisional conclusions

In the light of the above theoretical and practical/methodological analysis of the concept of validity and validation, it is clear that the new trends in language testing emphasise the importance of looking not only at test design but also at its actual and potential consequences. This encompasses the need to obtain empirical evidence on test validity from the test developer and other independent sources, to critically examine the extent to which the test can be used to draw inferences on future, context-specific language performance, and the correspondence between the test construct, content, and tasks and the target language use situation.

Test validation does not appear to be a 'one-off' event, a static moment in the testing process, but is rather a continuing process potentially investigating a great number of test aspects. These considerations lead to several implications in daily testing situations. According to the theory, a great amount of evidence should be gathered in order to validate a test and make sure that the inferences and consequential decisions we derive from a test are valid. There is not a single, exemplary testing situation to refer to in order to discuss the potential critical conditions of validation. Furthermore, tests have different impacts on the test takers according to the decisions which will be taken from test scores. This demonstrates that the process of validation might have a slightly different weight in different test situations thus involving an approach from a hands-on perspective. Time or money constraints are other elements playing a fundamental role in validation because in order to keep a continuing validating process in a test situation, it is indispensable to have human and material resources available to be used in this ongoing process.

The great deal of evidence to be gathered also implies the need to circumscribe the aspects of validity which are important in a given test situation. Nonetheless, two factors are essential in the analysis and setting up of the validation process: the test users and the context. Test users are here intended as both test takers and all the people who use the test to obtain information and make decisions. A preliminary overview of the test users' profile will help to understand what aspects should be investigated and what validation methods can be used. In high-stakes tests, where important decisions depend on both test takers and test users, it may be fundamental to validate the test in as many ways as possible in order to guarantee fairness. On the other hand, in an exit test after a language course, it might be useful to focus on content validity and construct validity to make sure the test is a good mirror of what the test taker has learnt.

As far as the context is concerned, we know that the testing situation deeply influences the validation process. The context includes the description of the language to be tested and the conditions under which it will be tested. There are a number of conditions that can potentially affect the testing process and these might include planning (presence or absence; time allowed etc.), complexity of input and/or expected output, gender or number of interlocutors, purpose of the test (source). It is important to identify those factors which may have a deep impact on the test process and which may thus be developed from the test specifications.

New views on validity are fundamental issues in language testing, but it is also clear that they represent a set of technical procedures that may be followed while developing a test according to specific needs. The provisional conclusion is that it is up to test developers to interpret the dynamic meaning ascribed to test scores by shifting from the 'self-serving role' of language tests to the coherent use of the 'multiple sources' of information that validation provides (Davies, Elder 2005, p. 811).

References

Alderson, J.C.; Clapham, C.; Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

American Educational Research Association; American Psychological Association; National Council on Measurement in Education (1966). *Standards for educational and psychological testing*. Washington D.C.: American Psychological Association.

American Educational Research Association; American Psychological Association; National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington D.C.: American Psychological Association.

Bachman, L.F. (1990). *Fundamental considerations in language*. Oxford: Oxford University Press.

Bachman, L.F.; Palmer, A.S. (2005). *Language testing in practice: Designing and developing useful language tests*. Oxford etc.: Oxford University Press.

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

Campbell, D.T.; Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research*. Skokie (IL): Rand McNally.

Cohen, A.D. (1999). «Language learning strategies instruction and research». In: Cotteral, S.; Crabbe, D. (ed.), *Learner autonomy in language learning: Defining the field and effecting change*. Frankfurt am Main: Lang, pp. 61-68.

Cronbach, L.J.; Meehl, P.E. (1955). «Construct validity in psychological tests». *Psychological Bulletin*, 52, pp. 281-302.

Cumming, A.H.; Berwick, R. (1996). *Validation in language testing*. Clevedon (UK): Multilingual Matters.

Davies, A. (ed.) (1968). *Language testing symposium*. London: Oxford University Press. Partial It. transl. in: Amato, A. (a cura di), *Il testing nella didattica linguistica*. Roma: Bulzoni, 1974.

Davies, A.; Elder, C. (2005). «Validity and validation in language testing». In: Hinkel, E. (ed.), *Handbook of research in second language teaching and learning*. Long Beach: Lawrence Erlbaum Associates, pp. 795-845.

Davies, A.; Elder, C. (ed.) (2005). *Handbook of applied linguistics*. Oxford: Basil Blackwell.

Fulcher, G. (1999). «Ethics in language testing». *TAE SIG Newsletter*, 1 (1), pp. 1-4.

Gipps, C.V. (1994). *Beyond testing: Towards a theory of educational assessment*. London: Routledge.

Goodwin, L.D.; Leech, N.L. (2003). «The meaning of validity in the new standards for educational and psychological testing: Implication for measurement courses». *Measurement and Evaluation in Counseling and Development*, 36 (3).

Hamp-Lyons, L.; Lynch, B.K. (1998). «Perspectives on validity: A historical analysis of language testing conferences abstracts». In: Kunnan, A.J. (ed.), *Validation in language assessment: Selected papers from the 17th Language testing research colloquium*. Long Beach: Lawrence Erlbaum Associates.

Hughes, A. (1996). *Testing for language teachers*. Cambridge: Cambridge University Press.

Kane, S.; Crooks, T.; Cohen, A. (1999). «Validation measures of performance». *Educational Measurement*, 18 (2), pp. 5-17.

Kunnan, A.J. (1998). *Validation in language assessment: Selected papers from the 17th Language testing research colloquium*. Long Beach: Lawrence Erlbaum Associates.

Kunnan, A.J. (2005). «Language assessment from a wider context». In: *Handbook of research in second language teaching and learning*. London: Routledge.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.

Lynch, B.K. (2003). *Language assessment and programme evaluation*. Edinburgh: Edinburgh University Press.

McNamara, T.F. (2004). «Language testing». In: Davies, A.; Elder, C. (ed.), *Handbook of applied linguistics*. Oxford: Basil Blackwell, pp. 763-783.

Messick, S. (1989). «Validity». In: Linn, R.L. (ed.), *Educational measurement*. 3rd edn. New York: Macmillan, pp. 13-104.

Messick, S. (1996). «Validity and washback in language testing». *Language Testing*, 13 (3), pp. 241-256.

Shaw, S.D.; Weir, C.J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.

Weir, C.J. (2005). *Language testing and validation: An evidence-based approach*. Houndgrave (UK): Palgrave-Macmillan.

Xi, X. (2008). «Methods of test validation». In: Shohamy, E.; Hornberger, N.H. (ed.), *Encyclopedia of language and education*. 2nd edn. Vol. 7: *Language testing and assessment*. New York: Springer Science + Business Media LLC.