A CONCEPT OF VALIDITY

To Thi Thu Huong[*]

Validity, a central concept in testing in general, has been a central concern of language testing (Anastasi 1988; Angoff 1988; Baker 1989; Hughes 1989; Messick 1989; Davies 1990; Bachman 1990; Alderson et al., 1995; Bachman and Palmer 1996). Traditionally, test validity is defined as "the fidelity with which it measures what it purports to measure" (Garett, 1947: 394 cited in Angoff, 1988: 19). The traditional view considered validity as a quality of the measuring instrument. In this view, language test validity is commonly deemed to consist of five different types of validity, defined by Morrow (1981: 13, emphasis added) as follows:

Face: The test looks like a good one [*in the eyes of lay people*].

Content: The test accurately reflects the syllabus on which it is based.

Predictive: The test accurately predicts performance [i.e. is *indicative of the same construct*] in some subsequent situation.

Concurrent: The test gives similar results to existing test [i.e. *measures the same construct*] which have already been validated.

Construct: The test reflects accurately the principles of a valid theory of foreign language learning.

Morrow's definition of predictive validity does not clarify whether the kind of performance the test should predict is a language performance, or another performance involving both language and non-language factors.

Dissatisfied with face validity, which "is the mere appearance of validity to the metrically naïve observer" (Stevenson 1985b:111), Weiner and Braun (1988) reduced the number to the "troika" of content validity, criterion validity (consisting of concurrent and predictive validity) and construct validity.

The division of validity into different types led to controversy on the importance of these types (Morrow 1981; Savignon 1983; Stevenson 1985a, 1985b; Anastasi 1988; Messick 1989; Davies 1990). Communicative theorists argued that content, face, and possibly predictive validity were the most important types (Morrow 1981; Hughes 1989). Supporters of psychometrics (Loevinger 1957; Messick 1975; Tenopyr 1977; Guion 1977, all cited in Angoff 1988: 28; Savignon 1983; Wood 1991) claimed that only concurrent and construct validity were worth considering in test validation, which is the process of collecting different kinds of evidence to support the interpretation and use of test scores for a particular purpose in order to establish a test's validity.

The recent trend in language testing discussions is to consider validity as a unitary concept with different types of validity as different aspects of validity (Messick 1989; Bachman 1990; Wood 199; Alderson et al 1995; Bachman & Palmer 1996; McNamara 1996). Within the new perception, construct validity is at the centre (Messick 1989) and is enriched with two new aspects of validity: response and consequential or washback

[*] Dr., Department of English and American Language and Culture, College of Foreign Languages - VNU

validity (Bachman & Palmer 1996: 29-35; McNamara 1996: 22-23). Response validity gives "information on how an individual responds to test items" (Alderson et al. 1995:176).

Consequential validity is "the potential social consequences of the proposed [test] use and of the actual consequences of the applied testing" (Messick 1989: 89). In language testing, consequential validity, subsuming washback (defined as the effects of assessment instruments on educational practices and beliefs (Cohen 1994: 41)) as one of its aspects, is the impact of language teaching, learning and curriculum, on "the teaching materials, the life chances of test candidates" or "other interested stake holders" (McNamara 1996: 23). Bachman and Palmer (1996) preferred to refer to consequential validity under the heading of "impact". Bachman and Palmer (1996) conceptualised the impact of test use as operating at both micro and macro levels. At the micro level, individuals are affected by a particular test use. The individuals include test takers, test users, decision makers using test scores, teachers, test takers' friends and relatives and future classmates, etc. At the macro level, the society and the educational system are affected. Thus, taking a systematic view, "virtually every member of the system is indirectly affected by the use of the test" (Bachman & Palmer 1996: 31). In this respect, consequential validity is much broader than washback validity, which often takes into account mainly test takers and teachers.

Over the years, validity has evolved from the concept of "test quality" to the concept of the use, the interpretation or the inferences made from test scores (Henning 1987; Anastasi 1988; Angoff 1988; Messick 1989, 1996; Alderson et al. 1995; Bachman 1990; Bachman & Palmer 1996). Messick explained the reasons for this change as follows:

> In general, content and criterion-related evidence, being contributory to score interpretation, are subsumed under the rubric of construct-related evidence. Yet, considerations of specific content and selected criteria resurface, in addition to the general construct validity of score meaning, whenever the test is used for a particular applied purpose. In justifying test use dividing validity evidence into three categories that are then merged into one, does not illuminate these nuances in the roles of specific content and criterion - related evidence as adjuncts to construct validity. What is needed is a way of dividing and combining validity evidence that forestalls undue reliance on selected form of evidence, that highlights the important though subsidiary role of specific content and criterion - related evidence in support of construct validity in testing applications, and that formally includes consideration of value implications and social consequences into the validity framework. (Messick 1989:20)

Different "types" of validity are now considered as different "methods" of assessing validity: "the more different "types" of validity that can be established, the better, and the more evidence that can be gathered for any one "type" of validity the better" (Alderson et al. 1995:171).

REFERENCES

1. Anastasi, A., *Psychological testing*, Sixth edition, New York, Macmillan, 1988.
2. Angoff, W.H. (1988) Validity: an evolving concept. In Wainer, H. & Braun H.I. (eds) *Test validity*. Hillsdale, New Jersey, Lawrence Erlbaum.
3. Bachman, L., *Fundamental considerations in language testing*, Oxford University Press, 1990.

4. Bachman, L.F & Palmer, A., *Language testing in practice: Designing and developing useful language test*. Oxford, Oxford University Press, 1996.
5. Davies, A., *Principles of language testing*, Cambridge, Basil Blackwell, 1990.
6. Linn, R.L. (ed)., *Educational measurement*, Third edition. New York, Macmillan, 1989.
7. McNamara, T., *Measuring second language performance*. London, Longman, 1996.
8. Messick, S. A., Validity. In Linn, R.L.(ed) *Educational Measurement* (3rd ed.). New York, Macmillan, 1989.
9. Messick, S.A., *The once and future issues of validity*, Assessing the meaning and consequences of measurement, in Wainer, H., Braun, H.I., (eds), *Test Validity*, Hillsdale, New Jersey, Lawrence Erlbaum, 1988.
10. Stevenson, D.K., (1985b) *Pop validity and performance testing*, In Lee et al, 111-118, 1985.
11. Wainer, H., Braun, H.I., (eds) *Test validity*, Hillsdale, New Jersey, Lawrence Erlbaum, 1988.

TẠP CHÍ KHOA HỌC ĐHQGHN, NGOẠI NGỮ, T.XIX, Số 4, 2003

MỘT CÁCH HIỂU VỀ ĐỘ GIÁ TRỊ

TS. Tô Thị Thu Hương

Khoa Ngôn ngữ & Văn hóa Anh - Mỹ
Trường Đại học Ngoại ngữ, ĐHQG Hà Nội

Bài viết này mô tả khái niệm về độ giá trị trong kiểm tra đánh giá nói chung, kiểm tra đánh giá năng lực ngôn ngữ nói riêng. Trước đây độ giá trị của một bài kiểm tra được quan niệm một cách truyền thống là độ trung thực mà bài đó đo được cái nó cần đo với nhiều loại độ giá trị khác nhau như giá trị bề mặt, nội dung, tiên đoán, tương đương và khái niệm. Xu hướng hiện nay trong kiểm tra đánh giá năng lực ngôn ngữ coi độ giá trị là một khái niệm đồng nhất với các loại độ giá trị truyền thống là các khía cạnh khác nhau của độ giá trị.