Journal of Education and Practice www.iiste.org
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol.8, No.18, 2017

A Snap Shot on Construct Validity in Language Testing: Definitions and Implications

Hamid Reza Babaee
M.A. in Teaching English as a Foreign Language, Faculty of Human Science and Literature
University of Guilan, I, R. Iran

Abstract

This paper presents an overview of construct validity and its' definitions and implications in language testing. Construct validity has gained momentum among language testers and teachers for the last decades or so. Unfortunately, not many people fully understand how to conduct a well designed construct validated test. It is hoped that this paper can provide some guidance and support to those who want to embark on construct validity investigations. Topics analyzed in this study included; the concept of validity and construct validity, qualitative and quantitative measures of construct validity; and the review of some studies that applied these measures.

Keywords: Language testing, Validity, Construct validity, Qualitative measures, Quantitative measures.

1. Introduction

As one of today's most extensively employed analytical tools, validity has been utilized prolifically in a vast majority of research paradigms in educational and other contexts and the paramount importance of it is obvious for everyone in these contexts, especially in language testing realms because of its determining effect on test takers' ability and their performances on tests. In these regard, the importance of validity has gone beyond the other psychometric measures and gets its' importance more than the reliability and other measures, as Fulcher (2010) put it "The codes and guidelines all place the concept of *validity* at the center of the testing enterprise. It is the concept of validity that guides our work in testing and assessment (p.19). Until 1989 the researchers in this domain took account of validity in the same way and defined it similar to each other (Fulcher, 2010). Validity as the name denotes refered to the extent to which the test measures what is supposed to measure as Hughes (1989) defined it "the extent to which the test measures accurately what it is intended to measure (p.22) or in a similar token, Garrett (1947) delineated it as "the fidelity with which it measures what it purports to measure" (p.394).

Beside the definitions, some types of validity have also been proposed by the researchers in this domain. One of these types is construct validity that has been thought to play as a foundation for the other kinds of validity. Bachman (1990) defined construct validity as "the extent to which performance on tests is consistent with predictions that we make on the basis of a theory of abilities, or constructs" (p. 255). From the advent of this type of validity into the educational contexts, a number of measures have been designed for the investigation of language tests from this perspective and some studies have tried to conduct the construct validity researches based on these measures. Furthermore, language test designers have began to utilize these measures for evaluating and increasing the validity of their test. However, among the people in different educational contexts, like EFL context, the significance of construct validity and the different ways of measuring it, have been underestimated and neglected. The purpose of this research is to shed light on this type of validity and make those people aware of the paramount importance, different aspects, and measures of the construct validity.

2. Validity

Over the years of introducing the validity to educational context its' definition has been encountered with major revisions and the other factors such as the ability of test takers have been taken into account. Indeed Messick (1989) was the researcher who introduced the new concept of validity as a unitary concept in educational, especially testing domain. He 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' (p. 13). This view of validity has also certified by Standards for Educational and Psychological Testing (1985 cited in Backman 1990):

> *Validity.... is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself (p. 9).*

In the similar view to validity, Cronbach (1971) emphasized that the instruments and other measurement devices are not the purpose of validation process, but, rather the inferences that are supposed to be drawn from the test

scores are of paramount importance and subject to validation. Furthermore, some years later, Messick (1996) reiterated his ideas about the unitary aspect of validity and went on to say that " In essence, test validation is empirical evaluation of the meaning and consequences of measurement, taking into account extraneous factors in the applied settings that might erode or promote the validity of local score interpretation and us"(p.246). Therefore the complex process of validation of tests requires the test constructors to take account of both the *evidence* that verify the interpretation **or** use and the *ethical principles* that provide the foundation or rationalization for that interpretation or use (Messick 1975,1980,1989). In consistence with Messick's view, Weir (2005) was also one of the researchers that defined validity based on test scores, he pointed out that "validity is perhaps better defined as the extent to which a test can be shown to produce data, i.e., test scores, which are an accurate representation of a candidate's level of language knowledge or skills. In this revision, validity resides in the scores on a particular administration of a test rather than in the test *per se*" (p.12). In another argument, Weir (2005) asserted that validity is a multifaceted concept and therefore, various types of complementary evidences are necessary to support any claims for the validity of scores on a test that lead to interpretation of the scores. He thus highlighted the existence of difference types of validity to providing adequate evidences that lead to the sound judgments based on score interpretation. Similarly, Bachman (1990) argued on the inclusiveness of validity and claimed that none of the validity types is complete and sufficient for sound judgment of test score leading to valid interpretation, and because of the relative variation of various kinds of evidences in different contexts, there is a need for the collection and interpretation of comprehensive information based on the evidences for establishment of validity of tests. Messick's (1989) work on validity inspired many scholars and change their understanding of this central issue in language testing. Fulcher (2010) was also one of those scholars who prioritize the notion of consequential validity over the other types. In this sense, she stated that validity "raises the question of the extent to which the score is relevant and useful to any decisions that might be made on the basis of scores, and whether the use of the test to make those decisions has positive consequences for test takers" (p.20). She also argued for five aspect of validity: substantive aspect, structural aspect, content aspect, generalizability aspect and external aspect. She claimed that substantive aspect of validity emphasizes the justification of inferences drawn from a test score about the knowledge, skills and abilities of a test taker. Structural aspect in her term is analogous with substantive aspect and investigates the test from the structural and scoring perspectives according to the skills and abilities of interest for the purpose of equipping the test to provide information on a number of different skills or abilities. Content aspect considers the correspondence between the content of the test and the content of a course of study, or of a particular domain of interest and emphasizes the congruence between them. The fourth aspect, according to Fulcher, inspects the generalizability of the test score beyond the activities and skills contained in the test. It probes answer for the question of whether a test is extrapolative of capabilities in contexts beyond those modeled in the test. Finally she declared that the external aspect highlights the convergence validity of test scores and seeks the relationship between scores on a test with scores of other measures of the same, or different, skills and abilities. Indeed, Filcher's five aspects of validity, takes account of validity from fundamental perspectives to test validation which eventually lead to evidence-based interpretation based on test scores that has construct validity as foundation for the interpretation of test use.

2.1 Construct validity

During the last decades, Literature has recommended different types of validity in the realm of language testing such as; criterion validity, construct validity, content validity, face validity and etc. The present study is concerned with investigating the construct validity of the IELTS Listening test. For the first time, the concept of construct validity is introduced by the American Psychological Association to deal with the adequacy of psychological tests (Cronbach, 1988). After that, the concept has gained its prominent importance in language testing for the purpose of interpretation of test scores and helped us in understanding validity as a unitary concept (Bachman, 1990). Various definitions have been proposed by researchers in this realm. Bachman (1990) defined construct validity as "the extent to which performance on tests is consistent with predictions that we make on the basis of a theory of abilities, or constructs" (p. 255). Messick (1975) claimed that construct validity is "a measure estimates how much of something an individual displays or possesses. The basic question of construct validation is, what the nature of that something?"(p.957). Messick (1980) further maintained that "Construct validity is indeed the unifying concept that integrates criterion and content considerations into a common framework for testing rational hypotheses about theoretically relevant relationships"(p. 1015). Fulcher and Devidson (2007) claimed that "Construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not 'operationally defined.' The problem faced by the investigator is, 'What constructs account for variance in test performance?"(p.182).Weir (2005) proposed theory based validity and context validity and related the construct validity to them and to the interactional competence and stated that construct validity was better characterized by an interaction of these two types of validity, not just by the individual abilities equipped by test takers. Carroll (1987) put forward the 'mental abilities' as constructs and

defined them in terms of mental tasks which students are thought to have in order to meet the demands of a test. Similarly, Fulcher (2010) defined constructs as "the abilities of the learner that we believe underlie their test performance, but which we cannot directly observe (p.96). In a similar vein, Cronbach and Meehl (1955) defined a construct as 'a postulated attribute of people, assumed to be reflected in test performance' (p. 283).

Messick (1996) categorized construct validity into six distinguishable aspects, in order to expound and clarify the key implicit issues in the concept of validity as a unified concept. These are generelizability, substantive, content, structural, external and consequential aspects. According to Messick (1996), the generelizability aspect of construct validity deals with the generelizablility of score interpretation beyond assessed task and to a broader context of the construct realm of interest. Substantive aspect highlights two significant notions; providing tasks that are representative of domain processes and content and providing empirical evidences that task processes are employed by test takers in responding to the task. In fact, substantive aspect of construct validity takes accounts of validity from content and rest takers perspectives in order to rationalize the evidences drawn from test scores. Content aspect of construct validity concerned with the representativeness and relevance of content and with technical content features (e.g. appropriate reading level, unambiguous phrasing and correct keying) - That is, the specification of the extent of the construct domain of interest to be assessed. Structural aspect relates the construct domain to the scoring criteria and rationalizes proposing the development of construct- based scoring criteria and rubrics. Indeed, this aspect of construct validity makes a relationship between the internal organization of construct realm and the internal organization of assessment (Messick, 1989). External aspect of construct validity deals with convergent and divergent verifications by exploiting multitrait-multimethod models of analysis (Campbell & Fiske, 1959). As Messick (1996), put it "external aspect refers to the extent to which the assessment scores' relationships with other measures and assessment behaviors reflect the expected high, low, and interactive relations implicit in the theory of the construct being assessed (p.251).Eventually, consequential aspect refers the intended and unintended impacts of interpretation of test scores and use on the process of teaching and learning. As a matter of fact, this aspect is concerned with issues such as bias and unfairness in test interpretation and use, positive and negative wash back in teaching and learning contexts. Indeed, these six aspects of construct validity proposed by Messick can be perceived as a comprehensive framework covering all multi-dimensional requirements indispensable for construct validation which lead to the sound and valid test use in related context.

2.1.2. Experimentations to investigate construct validity

In so far as construct validity is concerned, it has been measured by various means of experimentation. Cronbach and Meehl (1955) divided these measures into the five types: (1) Group differences; (2) Correlation matrices and factor analysis; (3) Studies of internal structure; (4) Studies of change over occasions; (5) Studies of process. In a similar vein, Messick(1989) referred to the five types of empirical evidences for construct validation as: (1) the examination of patterns of correlations with item scores and test scores, and among features of items and tests and scores on items and tests; (2) analyses and modeling of the processes underlying test performance; (3) studies of group differences; (4) studies of changes over time, or (5) investigation of the effects of experimental treatment. In fact, Messick's categorization is extracted and adopted from Cronbach and Meehl (1955) types of measures to construct validation with only some detailed differences. For example, in Messick's category the three types of measures namely; analyses and modeling of the processes underlying test performance, studies of group differences and studies of changes over time are exactly the same as Cronbach and Meehl's labeling. The difference is in incorporation of two measures of correlation matrices and factor analysis and Studies of internal structure of Cronbach and Meehl's labeling into one category of the examination of patterns of correlations among item scores and test scores, and between characteristics of items and tests and scores on items and tests and adding another category of investigation of the effects of experimental treatment.

According to Cronbach and Meehl (1955), the first type deals with testing directly the expectation that two groups function differently on the test based on a supposed construct. In this case, the required congruence between test and group designation should be coarse and augmentation of the congruence between the two would probably be an indicator of invalidity of the test. The second type is concerned with corelational investigations. The correlation between two tests that are supposed to assess the same construct is calculated. The more the correlation between two test in the same way, the more valid the assumption that the two test measure the same construct. The third type investigates the homogeneity of items within a test. That is, items are intercorelated with each other to support construct validity. Besides item-item correlation, item-test correlation is also examined using certain reliability formulas describing internal consistency. The fourth type considers constancy of test scores using related formulas such as; retest reliability and Cattell's N-technique. In this type, considering the high degree of stability as an indicator of construct validity of the test depends upon the theory defining the construct. The last type appraises the construct validity of a test through the observation of the person's process of performance. In this regard, for example, students' errors are investigated for the purpose of the evidence that whether the scores are related the supposed construct of interest or not. According to Bachman

(1990) two of the these types of construct validation measurements are important; correlational and experimental and one of them which deals with test takers' test performances is advantageous in a way that it provides new insights into factors effecting test performance.

Since the recognition of the construct validity as a significant psychometric measure affecting test interpretation and use, a number of perspectives have been reported as means for construct validation studies. Because of the fact that, there is a variety of measurement devices and strategies to construct validation, the more convincing strategies should be used for the purpose of gaining confidence related to the construct validity of the test that eventually lead to the sound judgments regarding test interpretation and use (Brown, 2000).

The strategies used in construct validation studies are divided into the two groups of quantitative and qualitative measures. The quantitative measures are employed in correlational studies such as; factor analysis, multi-trait/multimethod studies, the Structure equation modeling (SEM), item response theory such as; multidimensional Rasch model, the studies that deal with experimental design and the effect of the treatment on test scores and etc. The qualitative measures are concerned with the process of test takers' test taking performances, content analysis, and interview with teachers and lectures of the content domain of the interest and etc.

2.1.2.1. Quantitative measures

Two methods of quantitative construct validation named above; factor analysis and multi-trait/multimethod method, which are of prominent concern in construct validation studies are chosen to be delineated below.

Factor analysis

Factor analysis is a kind of statistical measure that is used extensively in the correlational studies of construct validation (Bachman, 1990). It aims "to represent a set of observed variables in terms of a smaller number of hypothetical variables' (Kim and Mueller, 1978a, p. 9). Factor analysis tries to discover fundamental and theoretical variables, or factors, that explicate the pattern of correlations within a set of observed variables (Farhady, 1983a; Oller & Hinofotis, 1980). According to Backman (1990), in construct validation studies, theoretical variables which are called 'factors', underlie the observed correlations and are concerned with constructs, test methods facets, and other impacts on test takers' performances in language tests and on the other hand observed variables are dealt with test scores or other measures. Finally, the process of factor analysis leads to factor loading that is an indication of the relationship among the test scores and the different factors identified after analysis. Two kinds of factor analysis have been recognized among the researchers in this realm; exploratory factor analysis and confirmatory factor analysis. According to Stevens (1996), in confirmatory factor analysis, a great deal of emphasis is put on strong theoretical and empirical underpinning that authorizes the researcher to originate a precise model to identify factor loadings and correlations. On the other hand exploratory factor analysis is utilized to discover data to find out the number or the features of factors that justify the covariation among variables when the researcher cannot form a premise about the number and the nature of the principal factors related to observed data.

One of the recent studies carried out on construct validation of listening comprehension tests exploiting factor analysis, had been that of Khoii and Paydarnia (2011). They investigated the construct validity of three different tests of EFL listening comprehension: multiple-choice, gap filling on summary and fill-in-the-blank. 91homogeneous EFL learners divided into three groups were invited to take the nine listening test, each of which was appeared in three formats. Having analyzed the data using statistical factor analysis method, they proved that multiple choice tests had the high construct validity more than the other formats. Furthermore, they used a repeated measure one-way ANOVA and revealed that, the fill-in-the blank items were the most problematic questions, while, the multiple choice items were proved to be the easiest for the test takers.

The multitrait-multimethod (MTMM) design

One of the other quantitative correlational methods to construct validation is the multitrait-multimetbod matrix that is originally introduced by Campbell and Fiske (1959) to the field of language testing. In this approach, According to Backman (1990) every measure contains a trait and a method. Therefore, tests ate are thought to be a combination of multiple traits with multiple methods. The advantage of this method to the other methods is that, the convergent and divergent validation of a test which are of prominent concern in construct validation is taken into consideration (Backman, 1990). Fulcher (2010) defined convergent validity as "the degree to which two or more independent measures of the same ability agree with each other" and the divergent validity as "the degree to which two or more measures of different abilities result in different patterns of scores" (p.320). In fact by considering convergent and divergent validation in multitrait-multimetbod matrix, construct validity of a test is investigated from various perspectives taking account of different traits and measures. In these regard, indispensable condition for construct validation is an establishment of the high positive correlation between two different measures assessing the same trait, for example, the high positive correlation between two tests of

cohesion and organization measuring textual competence and low or zero correlation between the two or the same methods measuring different traits, for example, the low or zero correlation between two tests of register and naturalness measuring sociolinguistic competence(Backman, 1990).

Since the introduction of the multitrait-multimethod (MTMM) design by Campbell and Fiske (1959), various studies have been administered to investigate the construct validity of language tests using this method. One of the studies was that of Pae (2012) who investigated the construct validity of the Pearson test of English Academic, utilizing multitrait multimethod approach. Utilizing this method, he investigated the validity of linguistic constructs and assessment method effects, and examined the convergent validity, divergent validity, and the effect of method variance in the field test of the Pearson Test of English Academic. In his study, Pae proposed three separate constructs and one combined-skill construct as traits that included; listening, reading, speaking, and integrated skills. He examined each construct by three different methods: integrated prescribed multiple-choice question format, constructed question format, and summarized question format. The participants of his study were adult English language learners (ELLs), whose age ranged from 17 to 59 years. The findings of the study proved the impact of the trait factors on ELLs' English performances and only partly impacts of the question format on their language achievement traits. The findings also showed that different constructs might be assessed by question type, confirming that some part of the variance was concerned with the three question-formats.

2.1.2.2. Qualitative measures

Despite the fact that quantitative methods have been thought to be the powerful means in experimentation of particular hypothesizes in construct validation studies, but they have been considered to have some limitations. These limitations lie in the essence that these measures are not fully capable of generating new hypothesis and more critically they take account of the product of test taking process(e.g. the scores) and neglect the processes underlying the test taking performances (Backman, 1990). Because of these criticisms, during the recent years qualitative measures have gained their paramount importance in construct validation studies. Amid these types of measures, the studies concerned with the process of test takers' test taking and content analysis will be explicated bellow.

Studies of test takers' processes underlying their test performance

As mentioned above, investigating the processes underlying the test takers' performances is one of the prominent methods in construct validation studies. The significant important of this method has been emphasized by Messick (1989) as follows:

> *In numerous applications of…techniques for studying the process, it became clear that different individuals performed the same task in different ways and that even the same individual might perform in a different manner across items or on different occasions... That is, individuals differ consistently in their strategies and styles of task performance.* (p.54)

There are a number of strategies for demonstrating this kind of construct validation studies such as; protocol analysis and computer modeling, the investigation of answer times and arithmetic modeling of the these times, the scrutiny of motives offered by test takers for choosing a specific response, and the examination of organized errors (Messick, 1989).

One of the researchers that used qualitative empirical research procedures to study the test taking processes in order to investigate the construct validity of language tests was Cohen (1984). He examined the types of strategies utilized by test takers during tests and their responses to different kinds of items and tests. He carried out his studies based on some of the strategies identified by Messick (1989) such as; verbal self-report data and found out a variety of strategies (e.g. guessing, using the immediate context, and translating) used by test takers during tests such as; cloze and multiple choice reading tests. He utilized the results of his studies as an evidence for construct validation of supposed tests.

Content analysis

Content analysis is one of the other methods exploited in construct validity investigation studies. In studies carried out by utilization of content analysis, taxonomic and analytical frameworks are employed to analyze the tasks contained in the tests. A variety of the studies have been carried out using these types of procedures. For example, More and Morton (2007) employed the framework named "classification scheme" with the other qualitative measure, that was interview with university staff, to investigate the correspondence between the writing section of the IELTS with the writing requirement of the university studies. They analyzed and compared the tasks from the two domains using a classification scheme developed for the study. They had developed their scheme using several types of recourses including previous survey studies of university writing, taxonomic and

analytical frameworks from discourse analysis and an initial analysis of their own data. In a task survey, they analyzed and compared a total of 155 academic assignment tasks and a corpse of IELTS task 2 items according to the four dimensions of difference of classification scheme: genre; information source, rhetorical function and object of enquiry. By analyzing the tasks in two domains, they discovered that there are some similarities and differences between IELTS writing and academic writing. The differences were due to the use of background knowledge and a limited range of rhetorical functions by IELTS test takers as opposed to the use of a variety of research-based processes and a diversity of rhetorical functions by academic students. They also found that the IELTS tasks were based on real world criteria (situations, practices and actions) as opposed to the academic tasks that were based on abstract entities (theories, ideas and methods). Having interviewed with the academic staff, they also confirmed the lecturers'' positive attitudes towards the nature of the IELTS task 2 format and the type of language instruction they imagined students would benefit in preparing for it.

In another similar study, Moore, Morton and price (2012) studied the relationship between the academic reading module of IELTS and reading requirements in academic context. They tried to investigate the construct validity of the IELTS academic reading module and its' congruence with the reading and general literacy requirements of university study. Their study was qualitative in nature utilizing two methods: the first method was a taxonomic and analytical framework adapted from Weir and Urquhart (1998) and their second method was an interview with university staff in the domain of academic reading in each specific reading discipline. Their adapted framework used to analyze the reading tasks in the two domains included two dimensions of difference: level of engagement and the type of the engagement. Level of the engagement was concerned with the level of the text with which the students need to engage in order to respond to the task (local vs global); on the other hand, type of the engagement referred to the way (or ways) by which students engage with the text in order to respond to the task, (literal vs interpretative). Having analyzed the reading tasks in two domains and interviewed with academic lecturers in each specific reading discipline, they found both similarities and differences between the reading requirements in the two domains. The similarity was due to 'local-literal' configuration in reading tasks of two domains that required from readers basic comprehension of relatively small textural units. The differences were due to different forms of engagement in academic purpose that required from readers critical evaluation of material both globally and interpretively. They also found differences in reading requirements across the specific reading discipline.

Overall, in spite of the divergence between the two types of construct validation measures, research has highlighted the combination of both quantitative and qualitative methods in these studies. For example, Bachman (1990) delineated this notion as follows:

> *In summary, the process of construct validation is a complex and continuous undertaking, involving both (1) theoretical, logical analysis leading to empirically testable hypotheses, and (2) a variety of appropriate approaches to empirical observation and analysis. It must consider the content relevance of the test, in terms of both the abilities measured and the method facets of the measurement procedure (p.270).*

3. Conclusion

Construct validity is a valuable tool for the evaluation of the language tests, when the purpose is fitting the abilities that the participants expected to have with the abilities that a language test is expected to test from those participants. The goal is taking the purpose of a test into consideration and making a balance between the purpose, the test and the social implications. In this sense, construct validity goes beyond its' realms and associates with social consequences and ethical considerations. Overall, since the recognition of construct validity as a prominent psychometric measure for validation of language tests use and interpretation, various definitions and measures have been proposed and employed in construct validation studies. Due to the complex process of construct validation requiring comprehensive and in depth methods, limitations related to the context of the study and other aspects may affect the ways of the study and hinder using some of qualitative and quantitative methods. Therefore, it should be mentioned that, because of the overriding influence of construct validity measures on designing language tests, taking the practicality of measures into consideration, test designers and researchers choose one of the methods or both of them and evaluate the construct validity of their test.

References

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

Brown, J. D. (2000). What is construct validity? JALT Testing & Evaluation SIG Newsletter 4(2), 8-12.

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

Carroll, J. B. (1987a.) 'New perspectives in the analysis of abilities' in R. R. Ronning, J. A. Glover, J. C. Conoley, and J. C. Witt (eds.): The Influence of Cognitive Psychology on Testing. Hillsdale, NJ: Lawrence Erlbaum Associates: 267-84.

Cohen, A. D. (1984). 'On taking tests: what the students report.' Language Testing 1, 1:70-81.

Cronbach, L. J. (1988). 'Construct validation after thirty years' in Robert L. Linn (ed.): Intelligence: Measurement, Theory, and Ptrblic Pobcy. Urbana, 111.: University of Illinois Press: 147-71.

Cronbach, L. J. (1971). 'Test validation.' In Thorndike, R. L. (ed.) Educational Measurement. Washington, DC: American Council on Education, 443–507.

Cronbach, L. J. and P. E. Meehl. (1955). 'Construct validity in psychological tests.' Psychological Bulletin 52,4:28 1-302.

Farhady, H. (1983a). On the plausibility of the unitary language proficiency factor. In J.W. Oller (Ed.), Issues in language testing research (pp.11-29). Rowley, Mass: Newbury House.

Fulcher, G. and Davidson, F. (2007). Language Testing and Assessment: An Advanced Resource Book. London and New York: Routledge.

Fulcher, G. and Davidson, F. (2007). Language Testing and Assessment: An Advanced Resource Book. London and New York: Routledge.

Garrett, H. E. (1947). Statistics in psychology and education. New York: Longman, Green.

Hughes, A. (1989). Testing for Language Teachers, 1st ed. Cambridge: Cambridge University Press.

Khoii, R. Paydarnia, S. (2011). Test Method Facet and the Construct Validity of Listening Comprehension Tests. The Journal of Applied Linguistics Vol. 4, Issue.1. 99-121.

Kim, J.-0. & Mueller. C., W. (1978a). Introduction to Factor Analysrs: What It Is and How To Do It. Beverly Hills, Calif.: Sage.

Messick, S. A. (1989). 'Validity' in Linn 1989: 13-103.

Messick, S. (1996). 'Validity and washback in language testing.' Language Testing 13, 241–256.

Messick, S. (1975). 'The standard problem: 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.

Weir, C. (2005). Language Testing and Validation. London: Palgrave Macmillan.

Moore, T and Morton, J. (2007). 'Authenticity in the IELTS Academic Module Writing Test: A comparative study of Task 2 items and university assignments', in IELTS collected papers: Research in speaking and writing assessment, eds L Taylor and P Falvey, Cambridge University Press, Cambridge, pp 197-248

Moore, T. Morton, J. Price, S. (2012) Construct validity in the IELTS Academic Reading test: A comparison of reading requirements in IELTS test items and in university study. IELTS Research Reports Volume 11. Swinburne University. IELTS Australia and British Council.

Oller, J. W. Jr. & Hinofotis. F., B. (1980). 'Two mutually exclusive hypotheses about second language ability: indivisible and partly divisible competence' in Oller and Perkins 1980: 13-23.

Stevens, J. (1996). Applied multivariate statistics for the social sciences (3rd Ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Pae, H. K. (2012). A psychometric measurement model for adult English language learners: Pearson Test of English Academic. Educational Research and Evaluation, 18, 211- 229.

Weir, CJ., and Urquhart, AH., (1998). Reading in a second language: Process, product and practice, Longman, New York.