

Investigating Administered Essay and Multiple-choice Tests in the English Department of Islamic Azad University, Hamedan Branch

Lotfollah Karimi¹ & Ali Gholami Mehrdad¹

¹ English Department, Hamedan Branch, Islamic Azad University, Hamedan, Iran

Correspondence: Lotfollah Karimi, English Department, Hamedan Branch, Islamic Azad University, Hamedan, Iran. E-mail: karimi.lotfollah@gmail.com

Received: June 8, 2012 Accepted: August 10, 2012 Online Published: August 22, 2012

doi:10.5539/hes.v2n3p69

URL: <http://dx.doi.org/10.5539/hes.v2n3p69>

Abstract

This study has attempted to investigate the administered written tests in the language department of Islamic Azad University of Hamedan, Iran from validity, practicality and reliability points of view. To this end two steps were taken. First, examining 112 tests, we knew that the face validity of 50 tests had been threatened, 9 tests lacked content validity, and 2 tests were impractical. Second, randomly the scores of 103 tests (8 multiple-choice and 95 essay-type), related to different academic courses, were fed into SPSS software to estimate their reliability statistically using Cronbach's alpha. The results showed that 7 multiple-choice tests had adequate reliability, 39 out of 95 essay-type tests had acceptable reliability, and totally there was a rather acceptable correlation coefficient among the scores of 95 essay-type tests. Finally, some suggestions were offered to improve the future tests.

Keywords: essay-type tests, multiple-choice tests, validity, reliability

1. Introduction

From the time when education began it has always been accompanied by testing. This relationship has been emphasized by many scholars and practitioners in both fields. According to Heaten (1988) as education has evolved new testing techniques have also emerged. On the same line Davies (1990) believes that language testing is central to language teaching.

To answer the questions such as "What type of test should we use?", "How long should the test be?", "How reliable should the test be?" and "Are our test scores valid for this use?" Bachman (1990) considers the specific uses for which the test is intended.

2. The Statement of the Problem

Testing and the future lives of testees are interwoven. Making tests, as measuring instruments of candidates' trait (a type of knowledge, say, vocabulary), is not an easy task and everybody who teaches is not necessarily a good test maker. It is naïve, for instance, to assume that one can develop valid tests of communicative language ability without reference to the construct which one is attempting to measure (Weir, 1990). To be able to make a reliable test, one should be aware of the rules and stages required. As a rule, for example, the test maker should know that the choices in a multiple-choice item should almost be of the same length, level of difficulty, and area of meaning. Writing a test is not a single shot. It should be passed through several stages such as describing the function of the test, planning, preparing items, revising the items, pretesting, and validating the test (Farhady, Jafarpoor, & Birjandi, 1994, 2006; Bachman & Palmer, 1996). The reliability of a test depends on many test method facets such as the facets of test rubric (test organization, time allocation, and instruction) and the facets of expected response (format, nature of language, and restriction on response) (Cohen, 1980; Bachman, 1990). Thus, making a good test is the business of those who have a hand in the field.

3. The Significance of the Study

Language testing and language teaching are so closely interrelated that it is virtually impossible to work in either field without taking the other into account (Heaten, 1988). Language testing provides goals for language teaching, and it monitors, for both teachers and learners, success in reaching those goals (Davies, 1990). It also leads to decision about placing students in the appropriate educational channel, educational quality, materials, and the tests themselves (Baker, 1989). Tests may abolish testees' rights to the extent that they are erroneous.

Revealing weaknesses in administered tests, as a result of such a kind of study, and providing the teachers, test makers, and researchers with feedback may help to bring about tests with high quality, facilitate the process of education, and take care of the testees' rights.

4. The Research Objectives

The main objective of the present study has been suggesting the teachers and/or test makers to be quite cautious while making a test and making themselves more knowledgeable about testing by studying the related literature. The second objective has been the improvement of testes being made by instructors in the future. Finally, facilitating the process of education, which is closely related to testing, has been taken into consideration.

5. Review of Literature

5.1 Reliability

According to Davies (1990) reliability is concerned with ensuring that a test is a measure (i.e. that whatever it is that it measures it does so accurately). It concerns the consistency of test judgment and result. It allows us to overcome doubts about the "mere" arithmetic of a test result. What reliability demonstrates is that there is consistency about the test as a whole, that scores from any section of the test are equivalent to scores on any other section (p. 5-6, 8, 21-6).

Baker (1989) has referred to "reliability" as "stability in the measure". In *inter-rater reliability* each candidate's performance is assessed independently by two assessors. By treating the two scores as results of two tests, the correlation coefficient can be calculated between the sets of scores. Any value less than 0.7 would be an indication of room for improvement in the briefing of assessors or system of scoring used (p. 60-61).

According to Valette (1977) test reliability refers to the constancy of the examination scores. Presumably if the same test were given to the same group of students, the performance of each student would show little variation. The requisites of dependable tests are the following:

Multiple samples: Many questions should be asked.

Standard tasks: All the students must be given the same item/items of equal difficulty.

Standard condition: All the students should take the examination under identical conditions.

Standard scoring: All tests must be scored in an identical manner (p. 44-6).

According to Mousavi (1999) reliability refers to consistency of measures across different conditions. If a student receives a low score on a test one day and high score on the same test later, the scores cannot be considered reliable indicators of the individual's ability. Similarly, if two raters give widely different ratings to the same sample, say those ratings are not reliable (p. 323-4).

A reliable test is one that produces essentially the same results consistently on different occasions when the conditions of the test remain the same. For example, for consistent results, we would expect the same amount of time to be allowed on each test administration. When a listening test is being administered, we need to make sure that the room is equally free of distracting noises on each occasion. If a guided oral interview were being administered on two occasions, reliability would probably be hampered if the teacher on the first occasion was warm and supportive and the teacher on the second occasion abrupt and unfriendly (Madson, 1983). The reliability of a test is a matter of how consistently it produces similar results on different occasion under similar circumstances (Oller, 1974).

Oller (1974) believes that if the test is administered to the same candidates on different occasions (with no language practice work taking place between these occasions), then, to the extent that it produces differing results, it is not reliable. Reliability measured in this way is referred to as *test/re-test* reliability. But the extent to which the same marks or grades are awarded if the same test papers are marked by (i) two or more different examiners or (ii) the same examiner on different occasions is termed *mark/re-mark* reliability (Heaton, 1988:162). The factors which affect reliability are the extent of the sample of material selected for testing, the administration of the test, test instruction (vague or clear), personal factors such as motivation and scoring the test (p. 163).

Harris (1969) defines reliability as the stability of test scores. To have confidence in a measuring instrument, we need to be assured that approximately the same result would be obtained (1) if we tested a group on Tuesday instead of Monday; (2) if we gave two parallel forms of the test to the same group on Monday and on Tuesday; (3) if we scored a particular test on Tuesday instead of Monday; (4) if two or more competent scorers scored the test independently (p. 14).

Summing up our discussion about reliability, we can conclude that consensus exists among testing scholars or practitioners on what reliability is, just their wordings are different. All of them believe that a test is reliable if its scores under different conditions approximately do not fluctuate.

5.2 Validity

The extent to which a test procedure is an adequate basis for decision-making is a question of its *validity* (Baker, 1989). Similarly, according to Mousavi(1999) validity is the extent to which the meaningful and appropriate inferences or decisions are made on the basis of test scores. Validity has also been defined as “the extent to which a test measures what it is supposed to measure (and nothing else)” (Heaton, 1988; Farhadi, Jafarpoor, & Birjandi, 1994).

Validity is divided into *face validity*, *content validity*, *construct validity*, and *empirical validity*. Here, only the first two have been dealt with.

5.2.1 Face Validity

If a test item looks right to other testers, teachers, moderators, and testees, it can be described as having face validity (Heaton, 1988). Face validity concerns the appeal of the test to the lay judgment, typically that of the candidate, the candidate’s family, members of the public and so on (Davies, 1990). Sometimes the students do not know what is being tested when they tackle a test. Sometimes they feel that the test doesn’t test what it is supposed to test. A test has face validity if it is carefully constructed, it has a [sic] well thought-out format, its items are clear and uncomplicated, its difficulty level is appropriate for students, and the condition for all students is the same (Mousavi, 1999).

5.2.2 Content Validity

Content validity may be defined as the extent to which a test measures a representative sample of the content to be tested at the intended level of learning. In other words, content validity refers to the degree of correspondence between the test content and the content of the materials to be tested (Farhadi, Jafarpoor & Birjandi,1994, 2006; Brown, 1996).

Content validity is a professional judgment, that of the teacher or testers. They rely on their knowledge of the language to judge to what extent the test provides a satisfactory sample of the syllabus, whether real (for achievement testing) or imagined (for proficiency testing) or of the theory or model (for aptitude testing) (Davis, 1990:23; Carmines & Zeller, 1991: 20). Similarly, Heaton (1988: 160) states that content validity depends on a careful analysis of language being tested and of the particular course objectives. The test should contain a representative sample of the course. On the same line Harris (1969) says:

If a test is designed to measure mastery of a specific skill or the content of a particular course study, ... the test should (my own emphasis) be based upon a careful analysis of the skill or outline of the course, and we ...expect the items to represent adequately each portion or outline(p. 19).

A test is said to be content valid if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned (Mousavi, 1999: 62). Other terms used for *content validity* are *curricular validity*, *course validity*, and, *text validity* (Ibid).

According to Oller (1979) content validity is related to the question of whether the test requires the examinee to perform tasks that are really the same as or fundamentally similar to the sorts of tasks one normally performs in exhibiting the skill or ability that the test purports to measure.

The criterion and the content models tend to be empirical-oriented while the construct model is inclined to be theoretical. Nevertheless, all models of validity require some form of interpretation: what is the test measuring? Can it measure what it intends to measure? In standard scientific inquiries, it is important to formulate an interpretive (theoretical) framework clearly and then to subject it to empirical challenges. In this sense, theoretical construct validation is considered functioning as a unified framework for validity (Kane, 2001).

Critically, Pedhazur and Schmelkin (1991) believe that “content validity” is not a type of validity at all because validity refers to inferences made about scores, not to assessment of the content of an instrument. The very definition of a construct implies a domain of content. There is no sharp distinction between test content and test construct.

5.3 Practicality

Another valuable quality of a good test found in the literature is *practicality*. By being *practical* it is meant that

the test should be useable within the limits of time and budget available. It should have a high degree of cost effectiveness (Oller, 1979).

Practicality is a matter of the extent to which the demands of the particular test specifications can be met within the limits of time and existing human and material resources Mousavi (1999).

On the same lines Harris (1969) refers to *economy and the ease of administration and scoring*. If a standard test is used, we must take into account the cost per copy. It should be determined whether several administrators and/or scorers will be needed, for the more personnel who must be involved in giving and scoring a test, the more costly the process becomes (p. 21-22).

Similarly, Farhadi, Jafarpoor & Birjandi (1994) state that *practicality* refers to the ease of administration and scoring of a test. In addition to these two major considerations, some other factors, such as the time of administration, cost of testing, ease of interpretation and application of scores, and availability of comparable forms contribute to the practicality of a test. Examining different scholar's view concerning reliability, validity and practicality, we come to the conclusion that all of them say, more or less, the same thing about these concepts although their wordings are different (159).

6. Method

6.1 Research Question

This investigation has been designed to answer the question "do the administered written tests possess validity, practicality and reliability as three major characteristics of a good test?"

6.2 Subjects

The subjects of this study were 112 tests (along with the scores obtained from them by the testees) which were randomly selected among many tests administered during some academic years.

6.3 Sources of Data and Instrumentation

The data were collected from three different sources. Those related to literature review were collected, via note-cards, from the sources introduced in the reference part. The second and third sources of data have been the tests (question sheets) and their scores respectively which were received from the authorities of the university through an official application letter.

6.4 Procedure

The purpose of this study was to determine the degree of validity, practicality and reliability of the administered written tests. To this end the researcher followed four phases. First, all the randomly selected tests were closely checked for their *face validity*. Throughout the study any weakness of any type was recorded. In the second phase, the content of each test was compared with the content of the syllabus of the related course to see whether they matched or not. In this way the degree of *content validity* of the tests was determined. In the third phase, all the tests were studied from *practicality* point of view, that is to say, the tests were evaluated to find whether they have been cost-effective or economical concerning time, budget, man power and instruments used to administer them or not. Finally, the scores of 103 multiple-choice and essay-type tests were fed into SPSS software system to be statically analyzed for their *reliability* using Cronbach's alpha and the findings are reported through the following tables.

The SPSS outputs were compared with 0.5 as the cut-point to see whether the results or scores of the tests were *reliable* or not. The reliability higher than 0.5 was considered as adequate and the one considerably below it as inadequate. Tables 1 through 8 below show the reliability of eight different multiple-choice tests administered and Table 9 shows the reliability/correlation coefficient of ninety five administered essay- type tests.

Table 1. The reliability of test A

Number of cases	Number of items	Calculated alpha
60	25	0.4259

As shown in Table 1, the reliability of test A is inadequate.

Table 2. The reliability of test B

Number of cases	Number of items	Calculated alpha
34	54	0.8041

As shown in Table 2, the reliability of test B is adequate.

Table 3. The reliability of test C

Number of cases	Number of items	Calculated alpha
44	36	0.7736

As shown in Table 3, the reliability of test C is adequate.

Table 4. The reliability of test D

Number of cases	Number of items	Calculated alpha
47	40	0.8001

As shown in Table 4, the reliability of test D is adequate.

Table 5. The reliability of test E

Number of cases	Number of items	Calculated alpha
35	40	0.6418

As shown in Table 5, the reliability of test E is adequate.

Table 6. The reliability of test F

Number of cases	Number of items	Calculated alpha
26	34	0.6515

As shown in Table 6, the reliability of test F is adequate.

Table 7. The reliability of test G

Number of cases	Number of items	Calculated alpha
22	27	0.6499

As shown in Table 7, the reliability of test G is adequate.

Table 8. The reliability of test H

Number of cases	Number of items	Calculated alpha
42	60	0.6360

As shown in Table 8, the reliability of test H is adequate.

Table 9. The reliability coefficient of 95 essay- type tests administered

Number of cases	Standardized item alpha	Calculated alpha
95	0.5667	0.4984

As shown in Table 9, the reliability/correlation coefficient of 95 essay-type tests may be considered adequate rounded to 0.5.

7. Principal Findings

- a) Findings related to face validity (found in 50 tests)
- 1) No suitable space between the lines of the texts
 - 2) Illegibility
 - 3) Not labeling the choices in multiple-choice tests by a, b, c...
 - 4) Ambiguous/misleading instructions
 - 5) Giving wrong address concerning the relationship between an antecedent and its reference word
 - 6) Using choices such as “all of the above” and “non of the above” for many times in multiple-choice tests
 - 7) Misspelling
 - 8) Using grammatical clues leading to the correct answer
 - 9) Subjectivity of some multiple-choice items
 - 10) The repetition of the same item in the same test
 - 11) Writing the number of items in Persian, but the item themselves in English
 - 12) Not highlighting the negative marker “not” in the instruction of multiple-choice items
 - 13) Low quality of duplicating/copying
 - 14) Writing different portions of the same test differently ,that is, with hand, computer and typewriter
 - 15) Not using correct or appropriate words in the instructions (“question” has been used instead of “item”; “in” has been used instead of “into”;...)
 - 16) Not taking care of capitalization
 - 17) Ignoring almost the equal length of choices in multiple-choice items
 - 18) Not putting the shared part of all choices in the stem/lead
 - 19) Using incorrect plural form
 - 20) Using wrong structure ,i.e., “ to be store” for “ to be stored”
 - 21) Jumbling /scrambling all parts of the test identification
 - 22) Not using any instruction for some items
 - 23) Forgetting to give or attach the text which has been required to be analyzed by the testees
 - 24) Using the phrase “all of them” for choice “a” in the case of some multiple –choice items
 - 25) Not using a hyphen , as affix marker, after a prefix and before a suffix
 - 26) Giving an incomplete item (two sentences, each belonging to a different language, have been asked to be compared, but only one of them has been given.)
 - 27) Using long instructions
 - 28) Asking testees to determine the number of a grammatical pattern instead of the structure of that pattern
 - 29) Not mentioning the time of answering the test
- b) Findings related to content validity

Nineteen tests, in one way or another, lacked content validity when compared with the demands of the syllabus.

c) Findings related to practicality

Two of the administered tests did not possess the characteristic of practicality, because they were time consuming as far as answering and correcting were concerned.

d) Findings related to reliability

- 1) Seven of multiple–choice tests (tests B, C, D, E, F, G, and H) had adequate reliability coefficient, for the calculated alpha of each was higher than the cut-point considered. However, one of them (A) lacked adequacy as far as reliability is concerned, because its calculated alpha was lower than the criterion (0.5 cut-point).

- 2) Totally there was a rather [my emphasis] adequate correlation coefficient among the scores of 95 essay-type tests ($\alpha=0.4982$ which can be rounded to 0.5). But, individually, the reliability of the scores of 39 out of 95 essay-type tests was higher than the cut-point considered (0.5), that of 43 was quite close to the cut-point so that one could round it, and that of the rest (13) was lower than it.

8. Discussion

The hierarchical purposes of this study have been:

- 1) Persuading the teachers and/or test makers to be more careful while making a test.
- 2) The improvement of the future tests , and
- 3) The improvement of the education.

To justify the basis of these purposes many written tests administered during some semesters in the English department of Islamic Azad University of Hamedan were evaluated. The results indicated that some tests suffered from practicality, validity, and reliability. This may be attributed to various types of factor. The first type is teacher and/or test makers. They may have made the tests hurriedly or have not been expert enough in test making. The second type include factors such as physiological and psychological conditions of testees, the structure of the test (i.e., homogeneity of items, speed, and length), administration, and scoring which, more or less and in one way or another, influence reliability. The third type consists of factors such as directions, difficulty level of the test, structure of items, arrangement of items and correct responses which contribute to validity or invalidity of the test (Bachman, 1990; Farhadi et al., 1994). As its limitation, this case study does not link its findings with those of the previous ones, if any. This may be attributed to the fact that such a study has not been conducted, at least, locally.

9. Suggestions

- 1) A workshop, where aspects of types of test are taught by experts in testing, is advised to be held for all those who teach regardless their field of study.
- 2) At the beginning of each semester, the syllabus of each course be given to the professors and they be asked to cover the contents of the syllabus during their teaching program.
- 3) Before being duplicated or copied, the tests should go through the steps of test making such as planning, writing items revising, pre-testing and assembling the final form.
- 4) When a single course is taught by different teachers and a common or shared test is administered to all the testees belonging to different classes, the test makers or professors should be as co-operative as required.

10. Recommendation for Further Research

Those who are interested in working in the areas related to such a study, may choose to work on Item Discrimination and Item Difficulty of the administered tests in any academic department or examination syndicates. The scholars of the other fields, rather than language, may replicate such a study concerning the tests administered in their own departments. Further, more investigation of the effect of washback of tests on teaching and learning is recommended, because this is a topic that is quite challenging and controversial.

References

- Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Baker, D. (1989). *Language testing: A Critical survey and practical guide*. Great Britain: British Library Cataloging in Publication Data.
- Brown, J. D. (1996). *Testing in language programs*. New Jersey: Prentice Hall Regents.
- Carmines, E. G., & Zeller, R. A. (1991). *Reliability and validity assessment*. Newbury Park: Sage publication.
- Cohen, A. D. (1980). *Testing language ability in the classroom*. Rowley, Mass: Newbury House.
- Davies, A. (1990). *Principles of language testing*. Oxford: Basil Blackwell Ltd.
- Farhady, H., Jafarpoor, A., & Birjandi, P. (1994, 2006). *Language skills testing: From theory to practice*. Tehran: The Center for Studying and Compiling University Books in Humanities (SAMT).
- Harris, D. (1969). *Testing English as a second language*. New York: McGraw-Hill, Inc.
- Heaton, J. B. (1988). *Writing English language tests*. London: Longman Group UK Limited.

- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319-342.
<http://dx.doi.org/10.1111/j.1745-3984.2001.tb01130.x>
- Mousavi, S. A. (1999). *A Dictionary of language testing*. Tehran: Rahnama Publications.
- Oller, J. W., Jr. (1979). *Language tests at school: A pragmatic approach*. London: Longman Group Ltd.
- Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associate.
- Valette, R. M. (1977). *Modern language testing*. New York: Harcourt Brace Javanovich, Inc.
- Weir, C. J. (1990). *Communicative language testing*. New York: Prentice Hall International (UK) Ltd.