ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 5, No. 2, pp. 471-475, March 2014

doi:10.4304/jltr.5.2.471-475

# The Process of Developing an Academic Reading Test and Evaluating Its Authenticity

Shahzad Karim
English Language Institute, King Abdulaziz University, Jeddah, Saudi Arabia;
Department of English, the Islamia University of Bahawalpur, Pakistan

Naushaba Haq
Department of English, the Islamia University of Bahawalpur, Pakistan

***Abstract*—The study is based upon a small project of developing an academic reading test in an English language teaching class and evaluating its authenticity. The article is basically in the form of a report explaining the whole procedure of developing the test, administering it with an English language teaching class and finally evaluating its authenticity. While evaluating the test the focus is on the validity and reliability of the test. The evaluation of validity and reliability has been carried out both quantitatively and qualitatively. The results show that the academic reading test was good and worked well in assessing the learners' reading ability. However, it cannot be called a perfect test and still needs some improvement. So finally some necessary suggestions have been provided for improving the test and making it more valid and reliable.**

***Index Terms*—academic reading test, authenticity, validity, reliability, evaluation**

## I. Introduction

The study is based on describing the process of developing an academic reading ability test undertaken as a class project and to evaluate its authenticity. The study is of multidimensional nature and consists of the following steps:

- Discussing the purpose of the test and the construct that underlies it.
- Describing the process of test development in relation to test validation.
- Evaluating the draft test on the basis of the results of the trial.

The basic purpose of the test is to develop the ability of designing and developing a language test of the students of a class of English language teaching. Thus, in this way along with the learning of the principles of language testing and assessment, the students have also been given a practical experience of designing and developing a test. Now we discuss the test in relation to the three steps mentioned above.

### *TEST SPECIFICATIONS*

A test's specifications tell us about what the test tests and how it tests it. "Test specifications are the blueprint to be followed by test and item writers, and they are also essential in the establishment of the test's construct validity." (Alderson et al., 1995, p. 9). The specifications of our test are:

**Purpose**

The test is based on assessing proficiency in reading skill which involves the ability of drawing meaning from a text and interpreting the information appropriately (Grabe and Stoller, 2002). The test was designed for the tertiary level learners of English enrolled in Foundation Certificate in English for Academic Purpose (FCertEAP) at English Language Academy (ELA), a language centre of the University of Auckland where various English language courses are conducted for foreign learners.

**Reading Texts**

The test is based on two reading texts.

1. TEXT 1 **"Temptation Free Television for Children?"** *has been adapted from: Dumont, P. (2001, September 28). Temptation-free television for children? Web World [UNESCO Communication and Information website]*.

2. TEXT 2 **"Cracking the Mysteries of Birds Migration"** *has been Adapted from issue 2666 of New Scientist magazine, 23 July 2008.*

**Text length**

Text 1 consists of about 1000 words whereas Text 2 consists of about 1100 words.

**Test Format**

The test is based on objective questions and consists of two sections having a total of 40 questions. Each question carries 1 mark.

Section 1 has18 questions and is based on text 1.

| | |
|---|---|
| Q1-5 | True/ False/ Not Given |
| Q 6-9 | Short answers in no more than three words |

| | |
|---|---|
| Q10-14 | Matching paragraphs with summarized statements |
| Q15-18 | Matching opinions with persons |

Section 2 has 22 questions and is based on text 2.

| | |
|---|---|
| Q 19-23 | Matching paragraphs with summarized statements |
| Q 24-27 | MCQs |
| Q 28-33 | True/ False/ Not Given |
| Q 34-40 | Matching terms with correct expressions |

**Time Allowed**

Time allowed for the test is 1 hour decided in consultation with ELA staff.

## II. Stages of Test Development

**Stage 1**

At first a comprehensive knowledge of language assessment and reading skill was provided by the teacher. Various topics like nature of reading ability, different skills involved in it, various types of reading test items and concepts like validity and reliability in tests were discussed in the class.

**Stage 2**

The two reading texts selected by the teacher were given to the students. They were divided into groups of 3 or 4 and each group was assigned to make a sample test of 20 questions/items consisting of at least 2 types of tasks.

**Stage 3**

Each group submitted his sample test which was analysed by the teachers. The sample tests were reviewed in the class by exchanging them among the groups in order to moderate each others tests in terms of format, item difficulty and rubrics clarity. Two more days were given to the groups to improve their tests.

**Stage 4**

One group took the responsibility of writing and formatting the final version of the test which was further reviewed in the class.

**Stage 5**

One group was given the responsibility of administering the test. The test was administered at ELA during the mid semester break. The participants i.e. the students enrolled in ELA module 3 & 4 of FCertEAP took the test and also filled a post-test questionnaire.

**Stage 6**

One group prepared the answer key and marked some papers. Two groups took the responsibility of marking the tests.

**Stage 7**

One group was given the responsibility of entering the data in SPSS software. The whole process of data entry and analysis was discussed in the class.

## III. Validity of the Test

The validity of a test can be judged by considering "does the test test what it is supposed to test?" (Alderson et al., 1995, p. 170). According to Hughes (2003) a test is said to be valid if it measures accurately what it is supposed to measure. Here we will discuss the validity of our test with reference to a **priori validity evidence** collected before the test event and a **posteriori validity evidence** generated after the test (Weir, 1993)

### *A. PRIORI VALIDITY EVIDENCE*

Priori validity concerns "what should be elicited by the test before its actual administration" (Urquhart & Weir, 1998). In priori validation we consider construct validity and context validity of the test. **Construct validity** refers to the extent to which a test score can be interpreted as a representative of the construct. The two major threats to construct validity are 'construct under-representation' and 'construct irrelevance' (Messick, 1995). **Content / Context validity** refers to the extent to which a test is considered representative of the real life conditions. Weir (1993) states that conscious efforts should be made to make the test representative of as many real life conditions as feasible. The two aspects of priori validity i.e. construct and content validity of our test may be discussed under the following headings.

**1. Response Format**

The selection of suitable task and response format is very essential for the true assessment of the skill. Firstly, we may discuss the selection of tasks in the test. If we analyse the tasks set in the test, we find that almost all the tasks represent the evaluation or assessment of macro-skills like skimming and scanning. The tasks mainly consist of following categories.

- True/False
- Short answers to questions
- Comprehension
- MCQs

Items 10-14 and 19-23 are based on matching the statements with the paragraphs of the texts from where they have been inferred. These items are based on comprehension and assess the skimming ability of the test taker as "skimming

is the process of rapid coverage of reading matter to determine its gist or main idea." (Brown, 2004, p. 213). Similarly the items 15-18 are based on matching each opinion with the appropriate person. These items are based on assessing scanning ability of the test taker as "scanning is a strategy used by all readers to find relevant information in a text." (Brown, 2004, p. 209)

Reading is a combination of two sub-skills 'micro-skills' and 'macro-skills'. Micro-skills involve processing letters, words, orthographic patterns, recognizing word classes like nouns, verbs etc. and understanding systems like tense and syntactic structures. On the other hand macro-skills are mainly concerned with the comprehension of semantic and pragmatic knowledge. Thus, considering the above mentioned description of the two skills, it is quite obvious that tasks set in the test are suitable for the assessment of macro-skills of reading. But they may not be considered appropriate for the assessment of micro-skills.

**2. Weighting**

"Weighting is concerned with the assignment of a different number of maximum points to a test item, task or component in order to change its relative contribution in relation to other parts of the same test." (Urquhart & Weir, 1998, p. 63). While considering the element of weighting in the test, it is quite obvious that some items have been given more weightage as compared to the others. For example there are ten items i.e. 1-5 and 28-33 in the form of true and false statements. Similarly, ten items i.e. 10-14 and 19-23 are also based on comprehension. While there are just four items 6-9 in the form of short answers to questions and just four items 24-27 are in the form of MCQs. Thus there is an unequal distribution of items and their score. Instead of this unequal distribution of the test items, equal distribution might have been carried out by adding some other tasks like answers to the questions in a sentence or two and summarising some part of the text. Though, it may be argued that the addition of these two items i.e. answers to the questions in a sentence or two and summarizing some part of the text may cause threat to the construct validity of the test as these two items also possess some element of writing skill. But the rationale is that all the four skills of language are integrated and depend on one another. As the assessment of speaking skill involves listening as well and without listening, speaking skill cannot be assessed, so in order to have a better and more authentic assessment of reading skill, some part of writing may be included.

**3. Test Rubrics**

Rubrics can be defined as: "the rubrics are the directions to the reader of what is required by the task. The way the prompt is worded can influence significantly what the candidate does i.e. what s/he perceives the purpose of the task to be." (Urquhart & Weir, 1998, p. 58). Rubric is a very vast term. It not only includes general instructions about the whole test but also the clarity of thought and expression in each question and statement requiring the candidate to perform a specific task or answer a question in a particular way. Urquhart & Weir (1998, p. 57) state: "the test rubric should be candidate friendly, intelligible, comprehensive, explicit, brief, simple and accessible. The rubric should not be more difficult than the text or task." So our reading test almost complies with the above mentioned qualities as the instructions in our test are vivid, written in simple words and sentences and grammatically correct. Further, the important points in each question have been bolded which also makes it easier for the candidate to understand the instructions completely and to give appropriate answers. Wherever required, 'Note' has also been mentioned with the instructions or statements of the questions.

**4. Order of Items**

It has been usually observed that some of the tests are a hotchpotch of items requiring recourse to different parts of the text in a seemingly random fashion (Urquhart & Weir, 1998). In a test of careful reading, the questions should follow a serial order as it decreases the difficulty level and increases the validity of the test. So, considering the sequence of items in the test, we find that most of the questions except a few have been set in the sequence of the text. However, scanning permits random access into the text. So, some questions which are not set in the sequence of the text are based on scanning ability.

The division of test into two sections, each based on a separate text, also makes it easier for the test taker to pay attention to one text at a time and, thus, enhances the validity of the test. Further, the division of each test into sections or paragraphs A, B, C, D, …… is also a good technique. The questions 10-14 and 19-23 are based on the comprehension of each section or paragraph rather than the whole text.

**5. Time Constraints**

In testing reading, the consideration of time constraints is very essential for the processing of text and answering the items because "if time allotment is not carefully planned, it may result in unpredictable performance (Urquhart & Weir, 1998, p. 65). Time allotted for the test is 60 minutes which is suitable as it has been reported by the candidates in the post-test questionnaire that they were able to complete their test within the specified time.

**6. Content Knowledge**

Urquhart & Weir (1998, p. 75) say, "The text should be suitable in terms of genre, rhetorical task(s) and pattern(s) of exposition and at an appropriate level of specificity and should not be biased or favour one section of the test population". The two texts selected for the test are appropriate to assess the reading skill as both the texts are of general interest and are not having any element of bias or favouritism to any section of the test population.

## IV. Statistical Analysis of Test Results

SPSS software has been used to analyse the results.

*A. The Measure of Central Tendency*

The central tendency is measured in the form of mean, mode and median. Their values are:

- Mean= 22.7
- Mode= 18, 21
- Median= 21

The low value of mean indicates the difficulty of the test; its high value indicates that the test is easy. The mean value of our reading test is 23/40, which indicates that the test was neither too difficult nor very easy. So, it proves the validity of the test.

*B. The Measure of Dispersion*

The two main features of dispersion are **range** and **standard deviation (SD)**. Range is obtained by subtracting the lowest score from the highest score. The highest score in the test is 37 and the lowest score is 11. So the range of score in our test is:

$$37\text{-}11= 26$$

SD is the average distance from the mean. The value of SD is 7.22. The bell shaped histogram shows that the test was neither too difficult nor very easy as scores are not contracted to one side; rather varying over a considerable range. Hence, it also proves the validity of the test.

*C. Reliability*

The reliability of a test is determined by the consistency of its scores as remarked by Hughes (2003, p. 36) "The more similar the scores would have been, the more reliable the test is said to be." The reliability of a test based on objective items is measured by three methods which are:

- The test-retest method
- The split-half
- The Cronbach's Alpha

We cannot determine the test-retest reliability of our test as it was conducted just once. However Cronbach's alpha and split-half methods are used to measure the reliability of our test. The value of reliability index (RI) for a completely reliable test is +1.0.

- In split-half method the reliability of a test is determined by dividing it into two sections. If the values of the two sections are closely related, the test will be more reliable. The value of RI for the two sections of our test is 0.994 which reveals that the test is highly reliable.
- The value of Cronbach's alpha for our test is 0.843 which proves the reliability of the test.

However, this high value of reliability of our test is due to the objective nature of the test.

**1. The inter-rater reliability**

In case of inter-rater reliability a value more than 0.8 is considered good. The value of inter-rater reliability in our test is 0.994 which is very high. It is also due to the objective nature of the test.

*D. Discrimination Index (DI)*

One of the major concerns of a test is to differentiate between weak and intelligent students and it depends upon the selection of test items. The more discriminating the items in a test are, the more reliable the test is. The value of the range of DI is from 1 to-1. If an item has a "0" value of DI, it shows that the item cannot discriminate between weak and intelligent student. A value of 0.4 for DI is considered good. In our test, we calculated the value of DI for about 8 items by selecting one from each task. The values are:

| Item1 | Item6 | Item11 | Item15 | Item20 | Item26 | Item31 | Item37 |
|---|---|---|---|---|---|---|---|
| 0.2 | 0.5 | 0.4 | 0.5 | 0.5 | 0.4 | 0.4 | 0.2 |

These values show that the test items are well discriminating. However, items with DI below 0.4 needs to be moderated.

*E. Facility Value (FV)*

FV indicates the level of difficulty of an item. The required of FV for a test item ranges from 0.33 to 0.67. A value below 0.33 reveals the difficultly of item, whereas a value above 0.67 shows that the item is easy. Some items having the value of FV below 0.33 and above 0.67 are shown in the table.

| item13 | item17 | item24 | item25 | item27 | item31 | item33 | item37 |
|---|---|---|---|---|---|---|---|
| 0.675 | 0.7 | 0.25 | 0.325 | 0.325 | 0.325 | 0.25 | 0.275 |

*F. Qualitative Data (Post-test Questionnaire)*

Qualitative data obtained through post-test questionnaire shows that a large proportion of students found that it was a fair test of their reading ability which is a positive comment about the content validity of the test. In response to the question about test tasks students found true/false the most difficult. It may be due to the addition of a third option i.e. 'Not Given' because sometimes it becomes difficult to distinguish between 'false' and 'not given'. However, the students found the text interesting but they had some difficulty in reading Text 2 as well as searching for the answers. It may be due to the reason that Text 1 is related to our daily life, whereas Text 2 is more specialized.

## V. Suggestions for Improvement

- The objective nature of response formats reduces the validity of the test. So, some other tasks like answers to the questions in a sentence or two and summarising some part of the text may be included to increase the validity of the test.
- Considering the weight of the two questions i.e. MCQs and True/ False, we find that there are eleven questions based on true/false and four questions based on MCQs making a total of 15 out of 40, the total score of the test. Thus, in a sense we can say that approximately 37% of the total of the test is based on questions having the chance of guessing which might affect the validity of the test. So, there should be an equal distribution of test tasks on the basis of their weight.
- The tasks set in the test are suitable for the assessment of macro-skills of reading. But they may not be considered appropriate for the assessment of micro-skills. So, some more tasks involving the assessment of microskills should be included in the test.
- FV of a test is not absolute. It changes with the change of participants. So, if the test is to be conducted with the same participants again, then the items with high FV (item 13 & 17) and low FV (item 24, 25, 27, 31, 33, 37) need to be moderated.
- Similarly items having the value of discriminating index below 0.4 need to be moderated.

## References

[1] Alderson, J. C. Clapham, C. & Wall, D. (1995). Language Test Construction and Evaluation. Cambridge: Cambridge University Press.
[2] Brown, H. D. (2004). Language Assessment: Principles and Classroom Practices. New York: Longman.
[3] Grabe, W. & Stoller, F. (2002). Teaching and Researching. Reading. Harlow: Longman.
[4] Hughes, A. (2003). Testing for Language Teachers. Cambridge: Cambridge University Press.
[5] Messick, S. (1995). Validity of Psychological Assessment. *American Psychologist*, 50 (9), 741-749.
[6] Messick, S. (1996). Validity and Washback in Language Testing. *Language Testing* 13 (3), 241-256.
[7] Urquhart, A. H & Weir, C. J. (1998). Reading in Second language: Process, product and practice. London: Longman.
[8] Weir, C. J. (1993). Understanding and Developing Language Tests. London: Princeton-Hall.

**Shahzad Karim** is a professional language teacher having nine years' experience of teaching English language and linguistics at university and college level. Currently, he is working as an English Language Instructor in King Abdul Aziz University Jeddah, Saudi Arabia. He is also working as an Assistant Professor in the Department of English, The Islamia University of Bahawalpur Pakistan. He has got the degrees of 'Master of Professional Studies - Language Teaching' from the University of Auckland New Zealand (2009), 'Master of Philosophy in Linguistics' (2008) and 'Master in English Language and Literature' (2003) from Pakistan.

He also possesses relish for research related to the field of Second Language Acquisition. He has got a couple of research papers published in national and international research journals. He is also the author of a book titled 'Implicit and Explicit knowledge and medium of instruction in Pakistan' published by Lambert Academic Publishing (LAP) Germany (2011).

Mr. Shahzad is also a Master Trainer involved in many teacher training programmes with the British Council, US-Embassy and Higher Education Commission of Pakistan.

**Naushaba Haq** is a professional language teacher having fifteen years' experience of teaching English language and linguistics at university and college level. Currently, she is working as a Lecturer in the Department of English, The Islamia University of Bahawalpur Pakistan. She has got the degrees of 'Master of Philosophy in Linguistics' and 'Master in English Language and Literature' from Pakistan.

She also possesses relish for research related to the field of testing and evaluation. She has got a couple of research papers published in national and international research journals.

Ms. Naushaba is also involved in many teacher training programmes with the British Council, US-Embassy and Higher Education Commission of Pakistan.