Ministry of Higher Education

And Scientific Research

University of Diyala

College of Basic Education

Department of English

Reliability and Validity in " Language Testing"

By

Roweda Abdul Khader Kamel

Supervised by

Nizar Hussein Waly

2014-2015

Dedication

To my Family , Teachers

and

Friends

Roweda

Contents

Introduction

Though some have argued whether testing is actually necessary at all, it is generally agreed that it is the most practical way to monitor and systematically rank students. And as test remain the most popular way to grade students fairly, the quality of their production would seem vital.

For test efficiency, validity and reliability need to be present. and as these two conditions are important for the effectiveness of testing, it is generally accepted that we can achieve a precise evaluation of our students if they are both consistent.

Unsurprisingly, however, the variables that exist in tests at times produce a range of results. This paper starts with an analysis of testing in general and of how the examination of validity and reliability is used as a means of quality control in test production. This is followed by an analysis of a listening test that is being used in a high school in Japan. Quantitative and qualitative results are analyzed to ascertain whether it is reliable and valid, and this followed by an evaluation of its overall effectiveness.

Section One

1.1 Language Assessment

This chapter explore show principles of language assessment can and should be applied to formal tests, but with the ultimate recognition that these principles also apply to assessments of all kinds.

In this chapter, these principles will be used to evaluate an existing, previously published, or create test. How do you know if a test is effective? For the most part, that question as :can it be given within appropriate administrative constraints? Is it dependable? Does it accurately measure what you want it to measure? These and other questions help to identify five cardinal criteria for "testing a test "practically, reliability, validity, authenticity, and wash-back .

We will look at each one, but with no priority order implied in the order of presentations. (H.Douglas Brown; 2004 :19)

1.2 Reliability

A reliable test is consistent and dependable. If you give the same test to the student or matched students on two different occasion, the test should yield the results. The issue of reliability of a test may best be addressed by considering number of factors that may contribute to the un reliability of a test. (H. Douglas Brown; 2004 :20)

1.3 Student -Related Reliability

The most common learner -related issue in reliability is caused by temporary illness fatigue a "bad day "anxiety, and other physical or psychological factors. which may make an "observed "score deviate from ones "true "score. Also included in this category are such factors as a test-takers "test wiseness "or strategies for efficient test taking. (H. Douglas Brown; 2004: 21)

1.4 Rater Reliability

Human error, subjectivity, and bias may enter into the scoring process. Inter -reliability occurs when two or more scorers yield in consistent scores of the same test, possibly for lack of attention to scoring criteria, inexperience inattention, or even preconceived biases.

In the story above about the placement test, the initial scoring plan for the dictation was found to be un reliable -that is, the two scorers were not applying the same standards.

Rater -reliability issues are not limited to contexts where two or more scores are involved .intra -rater reliability is a common occurrence for class room teacher because of un clear scoring criteria, fatigue bias toward particular "good "and "bad ".

Students or simple carelessness. When I am faced with up to 40 tests to grade in only a week, I know that the standards I apply -however subliminally-to the first few tests will be different from those I apply to the last few. I may be "easier "or "harder " on those first few papers or I may get tired and the result may be an in consistent evaluation across all tests. One solution to such intra -rater unreliability is to read through about half of the tests before rendering

any final scores or grades, then to recycle back through the whole set of tests to ensure an even -handed judgment. In tests of writing skills rater reliability is particularly hard to achieve since writing proficiency involves numerous traits that are difficult to define. The careful specification of an analytical scoring instrument however, can increase rater reliability.(H. Douglas Brown; 2004:21)

1.5 Test Administration Reliability

Unreliability may also result from the conditions in which the test is administered. I once witnessed the administration of a test of aural comprehension in which a tape recorder played items for comprehension, but because of street noise outside the building, students sitting next to windows could not hear the tape accurately. This was a clear case of unreliability caused by the condition of the test administration.

Other sources of un reliability are found in photocopying variations, the amount of light in different parts of the room, variations in temperature, and even the condition of desks and chairs. (H. Douglas Brown; 2004:21)

1.6 Internal and External Validity

Any research can be affected by different kinds of factors which, while extraneous to the concerns of the research, can invalidate the

findings. If terms are not consistently defined and used in the research the validity and the reliability of the results may be called in to question.

Findings can be said to be internally invalid because they may have been affected by factors than those though to have caused them or because the interpretation of the data by the researcher is not clearly supportable. They may be externally in valid because the findings cannot be extended or applied to contexts outside those in which the research took place. (Herbert W. Seliger and Elana Shohamy;1989: 95)

Factors affecting internal validity

Sometime the manner in which the research plan or experiment is conceived can affect the validity of the outcome. When the results of the research are deemed in valid because of the design or the manipulation of some of the internal components that make up the research, this is considered a problem of internal validity. We shall now discuss some of the major factors which can affect the internal validity of research.

1. subject variability
2. Size of subject population
3. Time allotted for data collection or the experimental treatment .
4. Comparability of subjects
5. History, attrition and maturation
6. Instrument/task sensitivity(Herbert W. Seliger and Elana Shohamy;1989: 95)

1.7 Content -Related Evidence

If a test actually samples the subject matter about which conclusion are to draw, and if it requires the test -taker to perform the behavior that is being measured, it can claim content -related evidence of validity, often popularly referred to content validity. You can usually identify content -related evidence observationally if you can clearly define the achievement to you are measuring.

A test of tennis competency that asks someone to run a 100 you dash obviously lacks content validity. If you are trying to assess a person's ability to speak a second language in a conversational setting. asking the learner to answer paper and pencil multiple -choice question requiring grammatical judgments does not achieve content validity.

A test that requires the learner actually to speak with in some sort of authentic context does . And if a course has perhaps ten objectives but only two are covered in a test then content validity suffers.(H. Douglas Brown; 2004:22)

1.8 Criterion-Related Evidence

A second from of evidence of the validity of atest may be found in what is called criterion -related evidence also referred to as criterion -related validity, or the extent to which the "criterion "of the test has actually been reached. It was noted that most class room -based assessment with teacher designed tests fits the concept of criterion -referenced assessment. In such test, specified class room objectives are measured and implied predetermined levels of performance are expected to be reached 80 percent is considered a minimal passing grade.

In the case of teacher -made class room assessment criterion -related evidence in best demonstrated through a comparison of results of an assessment with results of some other measure of the same criterion. For example, in a course unit whose objective is for students to be able to orally produce voiced and voiceless stops in all possible phonetic environments. The results of one teachers unit test might be compared with an independent

assessment -possibly a commercially produced test in a text book -of the same phonemic proficiency.

A classroom test designed to assess mastery of a point of grammar in communicative use will have criterion validity if test scores are corroborated either by observed subsequent behavior or by other communicative measures of the grammar point in question. (H. Douglas Brown;2004: 24)

1.9 Construct -Related Evidence

A third kind of evidence that can support validity but one that does not play a large a role for classroom teachers, is construct -related validity, commonly referred to as construct validity. A construct is any theory hypothesis, or model that attempts to explain observed phenomena in our universe of perceptions. Constructs may or may not be directly or empirically measured -their verification often requires inferential data. "proficiency "and "Communicative Competence " are linguistic constructs "Self-esteem "and "motivation "are psychological constructs virtually every issue in language learning and teaching involves theoretical constructs.

In the field of assessment, construt validity asks, "Does this test actually tap into the theoretical construct as it has been defined? "Test are, in a manner of speaking, operational definition of constructs in that they operationalize the entity that is being measured.

For most of the tests that you administer as a classroom teacher, a formal construct validation procedure may seem a daunting prospect. You will be tempted, perhaps to run a quick content check and be satisfied with the tests validity. But don't let the concept of construct validity scare you. An informal construct validation of the use of virtually every classroom test is both essential and feasible. (H. Douglas Brown; 2004:25)

1.10 Face Validity

An important faced of consequential validity is the extent to which "students view the assessment as fair, relevant, and useful for improving learning "or what is popularly known as face validity

" Face validity refers to the degree to which a test looksright and appears to measure the knowledge or abilities it claims to measure, based on the subjective judgment of the examinees who take it, the administrative personnel who decide on its use and other psychometrically Un sophisticated observers".

Sometimes students don't know what is being tested when they tackle a test. They may feel, for a variety of reasons, that a test isn't testing what it is "supposed "to test. Face validity means that the students perceive the test to be valid. (H. Douglas Brown; 2002: 96)

Section Two

2.1 Alternative Assessment procedures

Alternative assessment has been described as an alternative to standardized testing and all of the problems found with such testing. There is no single definition of alternative assessment. Rather a variety of labels has been used to distinguish it from traditional standardized testing.

Alternative assessment consists efforts that do not adhere to the traditional criteria of standardization, efficient, Co-effectiveness objectivity, and machine scorability.

Alternative assessment is different from traditional testing in that it actually ask students to show what they can do .Students are evaluated on what they integrate and produce rather than on what they are able to recall and reproduce. The main goal of alternative assessment is to "gather evidence about how students are approaching processing, and completing 'real -life 'tasks in a particular domain " .

Most important alternative assessment provides alternatives to traditional testing in that it

a. Does not intrude no regular classroom activities.
b. Reflects the curriculum that is actually being implemented in the classroom .
c. Provide information on the strengths and weaknesses of each individual student
d. Provide multiple indices that can be used to gauge student progress. (Jack C. Richards and willy A. Renandya ; 2002 :339)

2.2 Teaching for tests

One of the things that preoccupies test designers and teachers alike is what has been called the wash back or back wash effect. This refers to the fact that since teacher quite reason ably want their students to pass the tests and exams they are going to take their and teaching become dominated by the test and, especially by the items that are in it. Where non -exam teacher might use a range of different activities exam teachers suffering from the wash back effect might stick rigidly to exam -format activities. In such a situation, the format of the exam is deter mining the formal of the lesson .

Two points need to be taken into account when discussing the wash back effect, however. In the first place modern teat -especially the direct items included in them -are grounded for more in main stream classroom activities and methodologies than some earlier examples of the genre. In other words there are many direct test questions which would not look out of place in a modern lesson any way. But secondly, even if preparing students for a particular test format is a necessity." It is as important to build variety and fun into an exam course as it is to drive students towards the goal of passing their exam ". (Jeremy Hamer;2007:389)

2.3 Oral Interview

Prepare a 20-item guided oral interview appropriate for your student. In clued Yes /No, wh-, and either /or questions. Also include statement. In clued one or two questions that get the student to offer some kind of correction or modification. Also include at least one question requiring clarification.

Include at least two or three content questions. Make sure that most of the questions have some logical relationship to adjoining questions. (Harold S. Madsen;1983:177)

2.4 Scoring procedures

Indicate what your evaluation criteria are for the guided oral interview. Also prepare an objectified scoring system for the interview.

Administer the guided interview to at least five person (preferably your students).Using Your scoring system, calculate a numerical grade for each person and report these, finally, discuss any changes that you would recommend, based on your experience in administering and scoring the guided interview. (Harold S. Madmen;1983:177)

2.5 Evaluating Tests

A good evaluation of our tests can help us measure student skills more accurately. It also shows that we are concerned about those we teach -for example test analysis can help us remove weak items even before we record the results of the test.

This way we don't penalize students because of bad test questions. Students appreciate an extra effort like this, which show that we are concerned about the quality of our exams. And a better feeling toward our tests can improve class attitude, motivation, and even student performance.

Some insight comes almost intuitively. We feel good about a test if advanced students seem to score high and slower students to score low.

Sometimes students provide helpful "feedback "mentioning bad questions as well as questions on material not previously covered in class and unfamiliar types of test questions. Besides being on right level and covering material that has been discussed in class, good test are also valid and reliable. (Harold S. Madmen;1983:178)

2.6 Good Tests

Good tests are those that do the job they are designed to be and

which convince the people taking and marking them that they work. Good tests also have a positive rather than a negative effect on both students and teachers

A good test is valid. This means that it does what it says it will. In other words, if we say that a certain test is good measure of a student's reading ability then we need to be able to show that this is the case. There is another kind of validity, too in that when students and teachers see the test, they should think it looks like the real thing -that it has face validity. As they sit in front of their test paper or in front of the screen, the students need to have confidence that this test will work (even if they are nervous about their own abilities).

However reliable the test is face validity demands that the students think it is reliable and valid .

A good test should have marking reliability. Not only should it to fairy easy to make but any one marking it should come up with the same result as someone else .

However, since different people can (and do)mark differently, there will always be the danger that where tests involve anything other than computer scorable questions, different results will be given by different markers to minimize the effect of individual marking styles. (Jeremy Harmer; 2009: 167)

2.7 What is Validity ?

In the selection of any test, two questions must always be considered :

1. What precisely does the test measure?
2. How well does the test measure?

If the test is found to be based upon sound analysis of the skill or skills we wish to measure and if there is sufficient evidence that the test scores correlate fairly highly with actual ability in the skills area being tested, then we may feel reasonably safe in assuming that the test is valid for our purpose. (Harris;1969:19)

The validity of a test is the extent to which it is supposed to measure and nothing else. (Heaton;1990:159)

Validity is an integrative evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriate ness of inferences and actions based on test scores. (Bachman;1995:236)

Every test, whether it is a short informal class room test or a public examination should be as valid as the constructor can make it (Heaton;1990:159)

Test validity presupposes that the test writer can be explicit about what is to be tested and takes steps to ensure that the test reflects realistic use of the particular ability to be measured. As for as possible a test should limit itself to measuring only what is intended to test and not extraneous or un intended ability. (Weir;1993:19)

2.8 Empirical Validity

The best way to check on the actual effective ness of a test is to determine how test score are related to some in dependent ,outside criterion such as marks given at the end of a course or instructor's or supervisors ratings. If the evidence shows that there is a high correlation between test scores and a trust worthy external criterion We are justified in putting our confidence in the empirical validity of the test .

Empirical validity is of two general kind, predictive and concurrent or (status)validity depending on whether test scores are correlated with subsequent or concurrent criterion measures. For example if we use a test of English as a second language to screen university applicants and then correlate test scores with grades made at the end of the first semester, we are attempting to determine the predictive validity of the test. If on the other hand, we follow up the test immediately by having an English teacher rate each students English proficiency on the basis of his class performance during the first week and correlate the two measures, we are seeking to establish the concurrent validity of the test. (Harris;1990: 19-20)

2.9 Principles of Testing

In this section we shall present very briefly the major points of strength in foreign language testing. It should be remembered, however that no single test can have all these features, because each type of test has advantage and weaknesses. It is necessary, therefore, that a teacher should use more than one type of test, whether in the same examination or in a different one. (AI -Hamash and Abdul –Rahman;1977:197)

2.10 Subjective and Objective Tests

Subjective tests are those whose scoring is dependent on or affected by the prejudices or judgment of the examiner. Objective tests are those whose scoring is entirely dependent on the achievement of the testee uninfluenced by the personal feeling and prejudices of the examiner . To answer a subjective test, the testee has to use his own words and expressions where as to answer an objective test the testee has to select his answer from among four or even more alternative. (Darwesh and AL –Jarah: 1987: 8)

All test items no matter how they are devised require candidates to subjective judgment. for example, Candidates must think of what to say and then express their ideas as well as possible; in a multiple choice test they have to weigh up carefully all the alternative and select the best one .

Furthermore, all tests are constructed subjectively by the tester who decided which areas of language to test how to test those particular areas and what kind of items to use for this purpose.

Thus it is only the scoring of a test that can be described as being objective. This means a testee will score the same mark no matter which examiner mark's the test. since objective tests usually have only one correct answer for at least a limited number of correct answer, they can be scored mechanically. The fact the objective tests can be marked by computer is one important reason for their evident popularity among examining bodies responsible for testing large number of candidates. (Heaton;1990: 25)

Conclusion

As mentioned in the introduction this paper high stakes test are causing further demands to be met by test designers in creating tests that accurately measure what they are supposed to

Designers of tests must try to make their tests as possible. The validity and reliability of tests should be made available so there can be careful observation of how and what test are measuring. If the general consensus about a test is good, it can be considered as a bench mark for designers to work from. Though as mentioned in the back ground, as the pursuit of perfection is perhaps ultimately Un productive we can instead strive to encourage communication across administrators, designers and teachers to improve what we are ideally working to wards -more validity and reliability in tests and less invalidity and unreliability

☆ Bibliography ☆

- Bach man , lyle f.,(1995) , *fundamental Consideration in language teaching* , ox ford .

-brown , H . D , (2004*) language Assessment principles and Classroom practices , pearson* Education , new york .

- Dar wesh , Abdul jabbar and AL-jarah faris , (1987) *, an elementary course in testing English as a foreign language* , Baghdad , al-safad press .

- AL- Hamash , K . I . and Abdul-Rahim , s , (91977) , *Teaching English as a foreign language* , Baghdad .

- Harris , David , (1969) , *Testing English as a second language* , Mc Graw – hill , New York

- Harmer , J . ,(2009) , *the practice of English language teaching* , Iran .

- Harmer , J , (2007) , *how to teach English* , pearson , education limited , long man .

- Heaton , T.B , (1990), *writing English language tests* long man , new york .

- Madsen , Harolds .,(1983) , *Techniques in Testing* , London , oxford university press.

- Richards , Jack C. and Renandya , willy A . (2002), *Methodology in language Teaching* , Cambridge university .

- Seliger , Herbert W and shohamy , E., (1989), *second language research Method* , oxford

- Weir , Cyril , (1993) , *understanding and developing language Tests* , prentice – Hall , London .