==
Language in India www.languageinindia.com ISSN 1930-2940 Vol. 24:2 February 2024
==

Effect of Assessment (Validity and Reliability) of English Language Testing on (EFL) Students' Terminal Exam

Khatira Habibi, M.A. (TESOL)
Lecturer, English Department, Kabul University of Medical Sciences
khatera.habibi1@gmail.com

Meena Sadam, M.A. (TESOL)
Lecturer, English Department, Kabul University
habibilima88@gmail.com

==

Abstract

Background: Assessment is the systematic process of documenting and using empirical data to measure knowledge, skills, attitudes, and beliefs. In assessment the terms validity and reliability are the two most important elements in test designing and the process of scoring and administering of the test that directly influence learning outcomes. It is a very important issue for every instructor to be involved in test designing in order to measure what is intended to be measured, especially in EFL contexts.

Objectives: To find out the effect of validity and reliability of assessment in English as a Foreign Language (EFL) testing on students' final outcome's performance, and to explore the factors which directly influence the final testing results.

Method and Material: This is qualitative research; the research design is descriptive, and the data was analyzed descriptively. In this research, first we searched for the keywords which related to the research topic through the online academic journals and up to date sources such as (ERIC, ADRI). Out of 51 articles, we reviewed twenty-six of them which mostly explained the effects of validity and reliability of the test on students' performance. The articles were published between (1989-2022) years.

Results: The findings indicated that there is a significant link between the final testing result of students, and the English test we designed for English courses. The factors which influence the process of validity and reliability are shortage of congruence between the objectives of the curriculum, the format of the test, and teachers' inadequate understanding about the assessment.

Conclusion: In assessment the term validity and reliability are meaningful measurements that should be considered when attempting to evaluate the progress of students in any educational setting. It is pivotal issue for every instructor to be considered in test designing to assess what is intended to be measured especially for EFL instructors.

Keywords: Effect of Assessment, Test Designing, Validity and Reliability, Language Testing, fairness

Introduction

Assessment is the systematic process of documentation by using realistic data to measure performance knowledge, skills, attitudes, and beliefs. Assessment designers strive to create assessments that show a high degree of fidelity to the following traits (Content Validity, Reliability and Fairness).

Assessing students' knowledge and skills in the learning process requires valid and reliable assessment tools so that we can achieve the goal we set for specific courses. The assessment tools consist of the written test, oral test, field work, practical work, portfolio, conference, and the presentation. If we want to achieve the goal we set for specific educational field, we must develop a standardized test that assesses somehow the knowledge and skill of students through the whole course. Meanwhile, most of the students fail or get lower passing score in the final exam of English language which is a big problem in EFL context of language testing. Validity and reliability are essential topics in designing and administering and scoring a test.

Even exam atmosphere affects final result of any kind of testing to measure examinees' improvement. Reliability means if the test consistently measures what is supposed to be measured. Reliability is the degree to which an assessment tool produces stable and consistent results. The term ***validity*** means if the test measures what is supposed to be measured.

According to Hughes (1989) , validity and reliability are meaningful measurements that should be considered when attempting to evaluate the status of or progress of students in an education setting. Reliability means the test consistently measures what is supposed to be measured. Cozby, 2001 stated that "Reliability is the degree to which an assessment tool produces stable and consistent results". Cozy (2001, p.18). He emphasized that Reliability should be considered in three various stages such as before test, during the test and after the test. The term validity means if the test measures what is intended to be measured. It is a very essential issue for every instructor, especially EFL instructors to be considered in test designing in order to measure what is supposed to be measure. The term ***validity*** runs side by side with standard test. We call the test valid when the three focal components (content, objective, and test) relate to each other.

Teaching is a dynamic and flexible profession that requires changes. Most professional teachers bring changes in the supplementary and teaching methods, evaluation and testing method, even class atmosphere regularly every semester. They work for more professional and satisfactory manners of learning and teaching and to fulfill their students' needs.

Research Objectives: It is based on exploring the effect of validity and reliability of English language testing on (EFL) students' summative assessment performance and to find out the factors that influences the final consequences.

Research Questions

1. What factors influence the EFL students' final assessment results?
2. Can there be validity without reliability?
3. Why do most of the EFL students fail in the final exam however they enthusiastically take part in learning process?

Method and Material

This qualitative research was conducted through descriptive study design and the data was analyzed descriptively. In this research, we searched for the keywords which related to the research topic through the online scientific and up to date journals such as (ERIC, ADRI and etc).

Exclusion and Inclusion Criteria

We included articles that related to our topic. Additionally, we tried to find articles which were published in the reputed journals. The unreliable journals' articles were not included in this study.

After reviewing abstracts of all articles, we selected the ones which related to the EFL assessment. Out of 51 articles, we reviewed twenty- six articles that mostly explained the effect of assessment elements (validity and reliability) of the test on students' performance. Then, we summarized them and used the information in this research. Finally, we selected twenty-six of them which mostly focused on assessing students' performance in EFL context. The articles were published between (1989-2022).

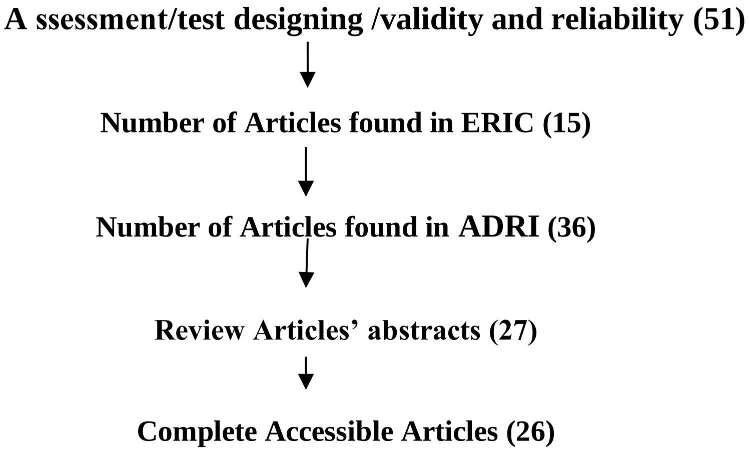
A ssessment/test designing /validity and reliability (51)
Number of Articles found in ERIC (15)
Number of Articles found in ADRI (36)
Review Articles' abstracts (27)
Complete Accessible Articles (26)

Figure (1): Shows the way of searching and selecting articles for the study.

Results

The findings revealed answers to the three research questions. One of the most significant discussions in language testing, which is becoming increasingly difficult to ignore has been the question of validity. In the past few decades, the conceptualization of validity has influenced extreme changes and has left its initial direction behind by focusing mainly on the question of whether the interpretations and actions based on the test scores are justified in the terms of evidential or consequential bases underlying test use (Messick, 2011).

The first section of the results explored the factors that influence the results of students' performance in terminal exam. The researchers stated their viewpoints on the impact of validity and reliability of the test and its role in assessment of English language testing on students' final result performance and the factors which directly influence the assessment process. The authors stated that validity should be considered 'the core of any form of assessment that is trustworthy and accurate.' Therefore, it has been borne in mind that validity, as an evolving complex concept, is closely related to the inferences made from assessment results.

The assessment element that helps this process of evaluation to be more valid is DIF (Differential Item Functioning). According to McNamara and Roever (2006), DIF originated in the early twentieth century and was used for the role of fairness in different tests to measure DIF. It was mainly prompted by researchers' interest in tapping social equity. The main purpose of DIF was to specify the confounding variables through purging items that highlighted the examinees' performance on tests. Mellenbergh (1989) defining item bias as conditional dependence, suggests that statistical tests and keys based on item response theory may be used for detecting biased items when items characteristic curves of the two groups being tested do not match. By relying on empirical or simulated data and combining information on the regression of item responses on hidden trait or observed test score and information on the hidden trait or observed test score distribution, it is possible to identify the biased items. Messick (1989) debated on the significant aspects of the tests in his validation framework-oriented testing research toward such conceptual variables as DIF, validity, and fairness causing a number of techniques for identifying biased items in different tests. As such, detecting differential functioning techniques turned into a primary concern in test development and test use whose main objective is to demonstrate that the interpretations and uses made of test scores are credible and trustworthy.

In the past few decades, Differential Item Functioning (DIF) has become increasingly an important area in language testing research. DIF is evidence of bias if the factor creating is not relevant to the construct characterizing the test. In short, if the factor is part of the construct, it is preferably called item impact instead of bias. In viewpoint of above remarks, most researchers have focused on

DIF and Differential Distractor Functioning (DDF) separately. However, the present study aims to critically examine the effect of hybridizing DIF and DDF to improve the validity and reliability of the language achievement tests. Moreover, the findings of this study will help test developers not only to become aware of some apparently invisible biases but to avoid them and subsequently to develop tests with much higher validity and great potential for fairly testing language skills of the examinees. Most researchers' articles address the integration of DIF and its possible effect on improving test validity and enhancing test fairness.

Therefore, this requires inference- based evaluative judgments that are reflective of truth and lead to specific interpretations and actions. According to Messick (2011, p. 5), "what is to be validated is not the test or observation device, but the inference derived from the test scores or other indicators". He also mentioned that basically because of the importance of precise and accurate inferences, test developers' accuracy in constructing tests has a great influence on the validity of assessment so that the suitability of the inferences made about the results of a test reflects the appropriateness of the conclusions derived from the testees' performance on the items including a specific test.

Consequently, test items are written to measure psychological attributes which are often not directly possible. In fact, they serve as representative measures of an unobservable psychological trait, a specific kind of knowledge, or psychomotor skill. Particularly, test items require examinees to employ their intellectual and thinking skills in order to answer the test items. This provides test developers with a physical measurement by which they can improve the validity of the test and the quality of the inference they make in order to judge the examinees' behaviors in terms of answering the test items or performing the required skills. (McNamara & Roever, 2006)

McNamara and Roever (2006) stated that test items act as stimuli whose main purpose is to prompt a prescribed or expected answer. The confirmation to a particular test item can be presentative of the fact that the examinee has acquired the intended characteristic or the attribute or has the ability to perform the skill taught. Since test are mainly applied for making high- stake decisions about the examinee, the assessment of the test result must be under careful examination and must be as fair as possible (Fulcher & Davidson, 2005; Shohamy, 2001; Stobart, 2005;Weir, 2005, as cited in

McNamara & Roever, 2006). Theoretically, biased test items may adversely affect test fairness and might have significant implication for policymakers, test developers, and test takers. Therefore, in developing a high-stake test, test developers should determine the extent to which a test item is affected by bias or impact.

The second section of results explored integration of validity and reliability in assessment. The item bias and item impact are closely tied to item validity and reliability that play a crucial role in language testing. As the most articles stated, item bias refers to the misspecification of the hidden ability space, where items measuring multiple abilities are scored as though they are measuring single ability. Moreover, according to Ackerman (2006), when two groups taking an identical or the same test possess different multidimensional ability distributions and the test items can possibly differentiate these levels of abilities on such multiple dimensions, then any unidimensional scoring method would unintentionally result in item bias. Therefore, item bias is an artifact of the testing procedure and is created when the source of the differential functioning of the item is irrelevant to the purpose of the test and the interpretation of the measures just because the item is tapping a factor which is over and beyond the targeted factor.

We agreed with the idea of the authors stated above. In any learning process there are many factors that directly influence the outcome. As we have different learning styles and strategy, the same as we have diverse socioeconomic, cultural, and educational backgrounds. A smaller number of EFL students have the chance of admission to a reliable and standard private language institutions/ course with well-educated instructors due to poor economic condition. Beside cultural issues are another factor, especially for Islamic countries (mixed classes of boys and girls) that limits the learning process and affects outcome.

In other studies which implemented in 2020 by Walker and Gocer Şahin, using differential item functioning, tried to evaluate interrater reliability as a guide to determine if two raters differ with respect to their rating on a polychromous rating scale or constructed response item. More specifically, they used differential item functioning (DIF) analyses to assess inter-rater reliability and compared it with traditional interrater reliability measures. The results showed that DIF procedures appear to be a promising alternative to assess the interrater reliability of constructed response items, or other polychromous types of items, such as rating scales.

On the other hand, item impact exists when one group of examinees tend to answer a particular test item more correctly than the other group of examinee because the two groups truly differ on the underlying ability (age, gender, intellectuality, skills, knowledge, learning style and learning strategy, environment, economy and etc.) In other words, item impact occurs when the item measures a relevant characteristic of the test without considering the actual differences existing between the two groups under assessment (Gelin, Carleton, Smith, &Zumbo, 2005). Clearly, the significant matters of test fairness and equity are essentially important because all examinees should enjoy equal opportunity to perform satisfactorily on a large –scale assessment and later being treated impartially in terms of their test scores (Moghadam & Nasirzadeh, 2020).

This portion of results investigated reasons of EFL students' failure in terminal exam. The distinction between item bias and item impact is defined and clarified by the purpose of the measure. Therefore, test developers should carefully analyze the test items to see that they are identified as presenting Differential Item Functioning (DIF). It is interesting to consider that DIF is not the direct indicator of bias in a test. Rather, as Karami (2011) maintained, that DIF is evidence of bias if the factor creating is not relevant to the construct characterizing the test.

A possible explanation for DIF is that it occurs when examinees from different groups with different demographic background like gender, ethnicity but the same true ability have a different probability of answering the item correctly. In short, if the factor is part of the construct, it is called item impact instead of bias. In viewpoint of above remarks, most researchers have focused on DIF and (DDF) separately. However, the present study aims to critically examine the effect of hybridizing DIF and DDF to improve the validity and reliability of the language achievement tests.

Discussion

As a result, test developers must make sure that the information obtained from such examinations was reliable and valid. This is only achievable if the items used in the test do not function differentially among different sub-population of examinees across different disciplines because of the factors which are not particularly relevant to the construct being measured. Most researchers stated that under identical testing conditions, it is expected that the examinees from different groups with comparable ability level show similar probability of responding correctly to a given item. Under such circumstances, DIF represents a modern psychometric method to the investigation of between group score variations. On the other hand, DDF is used to investigate the quality of a measure through understanding the biased responses across groups by shedding light on the potential sources of construct irrelevant variance by examining whether the differential selection of incorrect distractors attracts various groups differently (Penfield, 2010).

A possible explanation for DIF might be that it occurs when examinees from different groups with different demographic background like gender, ethnicity but the same true ability have a different probability of answering the item correctly. On the other hand, differential distractor functioning (DDF) is a phenomenon when different distractors, or inappropriate option choices, attract various groups with the same ability differentially. Martinkova and Drabinova (2018, p.505) suggested that when "a given item functions differently for two groups, it is potentially unfair, thus detection of DIF and DDF should be routine analysis when developing and validating educational and psychological tests". Moreover, the findings of this study will help test developers not only to become aware of some invisible biases but to avoid them and then to develop tests with much higher validity and great possible for fairly testing language skills of the examinees. Most researchers' articles address the integration of DIF and its possible effect on improving test validity and improving test fairness.

Conclusion

This paper has given an account of and the reasons for the importance of the test fairness by addressing the validity of designed general English language achievements test. The evidence of most research studies suggests that the social consequences of general English achievement tests and other language tests have been essentially important issue in recent decades. Instructors should bear in mind that careful design and development of any language test is a prerequisite and appropriate for any kind of assessment in EFL context. The current findings add substantially information to understanding of test validation and reliability and the crucial role it plays in the decision made about the test-takers, policymakers, and test developers' awareness about the importance of assessment. Validity and reliability are directly interrelated to each other in testing and evaluation.

Suggestions

We ask novice researchers to conduct quantitative research on the topic because there are many points that require exploration. We would like to recommend the head of any language learning private institutions to assign their members to establish their own exam committee to at least decrease the bias and distractors in assessment process. In recent years we have witnessed many exam problems before, during and after the exams contentiously that undergo the validity and reliability of assessment in every discipline.

===

References

1. Ackerman, T. A. (1992). A Didactic Explanation of Item Bias, Item Impact, And Item Validity from a multidimensional perspective. *Journal of Educational Measurement*, 29(1), 67–91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x
2. Adamson, B. (2003). *Assessment in Educational Setting* Retrieved from http://knol.google.com/k/micha-b-paradowski/assessment /2qpvzotrrhys1 /23#
3. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1989). *Standards for Educational +and Psychological Testing*. Washington, DC: Authors.
4. Bond, T. (2003). Validity and assessment: a Research Measurement Perspective. Metodoliga de las Ciencias del Comportamento, 5(2), 179–194.
5. Cozby, (2001). English Teaching and Assessing: Approaches and Procedures for Teaching Grammar. Practice and Critique. *Journal of language assessment*. 5 (1) 122-141 Retrieved from http://education.waikato.ac.nz/research/file/etpe/2006v5n1nar1.pdf
6. Cronbach, L. J. (1989). Test Validation. In R. L. Thorndike)Ed.). *Educational ,Measurement* (2nd ed.). Washington, D. C.: American Council on Education.
7. Dekeyser, R.M. (2000). The Robustness of the Critical Periods Effects in Second Language Acquisition. *Studies in Second Language Acquisition*.22 (4) Retrieved from http:// eltj.oxfordjournals.org/Kabul Education University
8. Fulcher, G., & Davidson, F. (2013). The Routledge Handbook of Language Testing. Routledge. https://doi.org/10.4324/9780203181287.

9. Jalili, T., Barati, H., & Moein Zadeh, A. (2020). Using Multiple-Variable Matching To Identify EFL Ecological Sources of Differential Item Functioning. *Journal of Teaching Language Skills*, 38(4), 1–42.
10. Karami, H. (2011). Detecting Gender Bias in A Language Proficiency Test. International Journal of Language Studies, 5(2), 27–38.
11. McNamara, T., & Roever, C. (2006). Psychometric Approaches to Fairness: Bias and DIF. The International Journal of Language Learning, 80(4), 808–820.
12. Mellenbergh, G. J. (1989). Item Bias and Item Response Theory. *International Journal of Educational Research*, 13(2), 127–143. https://doi.org/10.1016/0883-0355(89)90002-5.
13. Messick, S. (1990). Test validity and the ethics of assessment. American psychologist, 35(11), 1012–1027.https://doi.org/10.103 7/0003-066X.35.11.1012.
14. Messick, S. (2011). Meaning and values in test validation: The science and ethics of assessment. Educational researcher, 18(2), 5–11. https://doi.org/10.3102/0013189X018002005.
15. Moskal, B.M., & Leydens, J.A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation,* 7(10). [Available online: http://pareonline.net/getvn.asp?v=7&n=10].
16. Nitta, R., & Gardner, S. (2005). A new science of educational testing and assessment *in ELT course books:* ELT Journal Volume 59/1 January 2005©Oxford University Pressdoi:10.1093/elt/ccioo1
17. Pamela, A. (1994). Measurement and Evaluation. The University of Michigan Journal 23 (2), 5-12
18. Penfield, R. D. (2010). An Odds Ration Approach For Assessing Differential Distractor Functioning Effects Under The Nominal Response Model, *Journal of Educational Measure*, 45 (3), 247-269.
19. Scribbr, H. (2009). What’s The Difference between Reliability and Validity? retrived from :www.scribbr.com › frequently-asked-questions › reliability
20. Shohamy, E. (2001). Democratic Assessment as an Alternative. *Language Testing Journal, 18*(4), 373-391.
21. Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF Analysis of an L2 Vocabulary Test. Language Testing, 17(3), 323–340.https://doi.org/10.1177/026553220001700303.

22. Walker, C. M., & Gocer. Şahin, S. (2020). Using Differential Item Functioning to Test for Interrater Reliability in Constructed Response Items. Educational and Psychological Measurement, 80(4), 808–820.
23. Widodo, H.P. (2006). English Teaching and Assessing: Approaches and Procedures for Teaching Grammar. *Practice and Critique*. 5 (1) 122-141 Retrieved from http://education.waikato.ac.nz/research/file/etpe/2006v5n1nar1.pdf
24. Williams, J. D., Abt, G., & Kilding, A. E. (2010). Ball-Sport Endurance and Sprint Test (BEAST90): Validity and Reliability of A 90-Minute Soccer Performance Test. *Journal of Strength and Conditioning Research*, 24(12), 3209-3218. doi:10.1519/JSC.0b013e3181bac356
25. Wong, C. (2012). Examining the Effectiveness of Validity and Reliable Assessment. *Journal of English Assessment Tools,* 5 (1) 177-200.
26. Zumbo, B. D. (2020). A Handbook on the Theory and Methods of Differential Item Functioning (DIF), 5 (1) (pp. 1–57). National Defense Headquarters.

==

Khatira Habibi, M.A. (TESOL)
Lecturer, English Department, Kabul University of Medical Sciences
khatera.habibi1@gmail.com