Investigating the Quality of Language Tests; Calculating the Reliability Coefficient of Placement Test as a Case Study.

Jamal Ali Omar[1] - Sardar Abdulla Hussein[2]

[1+2] Department of English Language, College of Basic Education, University of Raparin, Rania, Kurdistan Region, Iraq.

Abstract

Language measurement tools should be valid and reliable by which the test scores obtained help drawing meaningful inferences and, fairly enough, significant decisions are made. This study attended to the reliability of an English language placement test. The aim of the study is to find out the reliability estimate of that type of test. Based on the Classical Test Theory, the reliability coefficient of the test should be computed to find out the range of the measurement error and achieve an estimate of the candidates' true scores. To this end, the data obtained from the administration of University of Raparin's Language and Development Center Placement Test[i], which was administered to 889 freshmen students, was analyzed. The method is descriptive and exploratory in which, by using the split half method, the three formulas – Kuder-Richardson 20 and Spearmen-Brown Prophecy – were used to calculate the reliability estimate of the test. The results showed that the reliability estimate of the test is 0.81. They also showed that the number of the items of the test should be doubled in order to achieve the desired reliability of 0.9 and above. Thus, based on the results obtained, it is recommended that the test structure should be improved to achieve a higher rate of reliability estimate. It is also recommended that the test should be modified; test items need to be increased and the content amended to cover all the language skills. Additionally, it is recommended that other sources of unreliability, namely, test administration and individual test items need to be studied to find the source of measurement error.

Keywords: Reliability Coefficient, Split-half Method, Placement Test.

Article Info:
DOI: 10.26750/Vol(10).No(1).Paper11
Received: 10-May-2022
Accepted: 27-June-2022
Published: 29-March-2023
Corresponding Author's E-mail:
jamal.ali@uor.edu.krd
sardar.mangori@uor.edu.krd

1.1 Introduction

The recent development in the field of language testing and assessment entails designing valid, reliable and practical tests. The focus on developing the featured language assessment constantly intensifies worldwide, and the need for knowledge of analysis of language test data is vastly turned into invaluable requirement (Aryadoust and Raquel, 2019). This research, as such, which attempts to calculate reliability coefficient, is an academic endeavor toward that direction which is genuine in its nature. Reliability, particularly as "an essential quality [and feature] of any MEASUREMENT PROCESS" (Mousavi, 2012, p. 621, emphasis in origin), is the degree of consistency of the test scores obtained by a measurement tool. It refers to the internal consistency and measurement errors of the measurement tool, that is 'the test' itself. The degree of measurement errors of language tests basically originates from three sources; the test construction, test administration and scoring related variables. Hence, investigating any of these areas is invaluable in that the scores will be the mere evidence by which inferences are drawn and consequently decisions are taken about test-takers.

The dominant research method in the field of language testing and assessment is the quantitative research type (Rahman, 2020). The information obtained from quantifying the variable and the statistical analysis of test data contributes not only to the development and improvement of language tests, but also to the extent to which the type of results obtained from them should be relied on or not. In other words, the tests' reliability needs to be scrutinized in order to avoid making decisions that may have unintended consequences. Furthermore, graphically and statistically describing and displaying test results will provide test developers and users with significant information on how the students performed on a test (Brown, 2005). The question generally raised is that should stakeholders and decision-makers fully trust the scores obtained by tests? To reduce the doubts and validate the results of a test, the best way is to calculate the reliability estimate of test scores, or to compute the reliability coefficient – a quantitative expression of reliability (Mousavi, 2012) - which is to mathematically quantify the reliability of test scores. Notably, the use of test data analysis results would primarily be crucial in developing reliable language tests. These results contribute to continuously revising and improving the measurement tool.

This research attempts to compute and investigate language test reliability coefficients. The study examines a placement test, which is developed and administered by the University of Raparin Language and Development Center (UoRLDC). A statistical analysis is conducted to the data obtained by the test annually administered to students who are newly admitted to the university. The analysis aims at analyzing the obtained data to identify any source of unreliability and measurement error by computing the test's reliability estimate, standard deviation and the standard measurement error. Computing these measures significantly contribute to confirming the degree of reliability of a language test. The rationale is that the targeted cohorts' English language ability is indicated for the stakeholders for further English language study as part of their academic and university degree requirements. Hence, a considerable high-stakes decision based on the results of the test will eventually incur as the test is used to place learners into suitable English language learning courses and classes. The main aim of the research is twofold. First, it is to use the test data analysis results that can contribute to improving and evaluating the test. Secondly, it is further to expose to the readers the degree of reliability of the placement test which is administered annually at University of Raparin as a general policy to enhance the English language ability improvement of the students. The overall plan is that transparent, annually revised and improved assessment procedures are considered to be a genuine academic requirement. As the case of this research is UoRLDC, a general overview about it is needed which is given below.

Original Article / Doi: 10.26750/Vol(10).No(1).Paper11

1.2 UoRLDC Placement Test

UoRLDC Placement Test has become a requirement to be administered to first-year students at University of Raparin. It is carried out annually to identify the new students' English language level, and it is part of the university's long-term plan and mission to improve students' English language ability. Based on the test results, students are divided and placed into beginner, elementary and pre-intermediate groups of EFL learners. Other sub-groups are also formed to the primary levels with the purpose of designing unique instructional plans to raise their language proficiency levels. The consequential decisions based on its results about the test-takers' upcoming language learning classes and courses are made. A brief description of the structure of the test is given below:

The test consists of three sections; section one is about candidates' biodata, section two comprises the grammar and vocabulary test items, and section three contains the reading comprehension test items. An outline of each of the sections is like; section one consists of two items by which the test-takers full name and department are elicited. Section two comprises 34 test items to assess the test-takers knowledge of grammar and vocabulary. The final section consists of a reading passage, which is followed by six items to assess their reading comprehension skill. Multiple choice questions is the technique used in the test, and hence, the students' answers were scored automatically by Google Form as the service includes the feature. The total number of testing items is 40, and the test is out of 40 points. Eight hundred eighty-nine first-year students from six colleges (14 departments) at University of Raparin sat for the test, which was administered electronically in computer labs. The administration environment was fully controlled to be the same for all the participants, as it may have impact on the test-takers' results. Each of the test-takers is given one hour to answer all the 40 items online. The answers were directly scored by the software created for that purpose. Thus, it is believed that the two other variables as the sources of unreliability were controlled by the test administers.

2.1 Research Background

Researchers, stakeholders and test designers have always been concerned about the consistency of the scores obtained by administrating language tests across time and test forms. Consistency of the results is essentially referred to as reliability, the extent to which the results obtained by a measurement tool truly represent the test-takers' real-time performance and ability. Reliability "is a central concern for interpreting assessment results, even to the point that it is an important part of most validity arguments" (Chapelle, 2013, p. 4918). According to Fulcher (2013), reliability is about the fluctuation of test scores (consistency of the results obtained by two different administrations of the same test, (Hughes, 2003). Based on Lado (1961), Fulcher (2013) maintains that the source of fluctuation of the test scores might be from the following variations; "variation in conditions of administration, the quality of the test itself and variability in scoring" (p.46). Likewise, according to Brown (2004), there are a few factors that need to be considered as they contribute to the unreliability of tests; these are fluctuations in students' performance, scoring, test administration, and in the test itself (p. 20-21). Thus, the source of unreliability has long become one of the areas to be studied in order to find out where the measurement error is. Indeed, reliability studies are much concerned about where the problems lie causing the scores not to be consistent.

Additionally, Mousavi (2012) defines reliability with three different approaches, which are summarized in three related expressive terms; stability, accuracy and error of measurement (p. 622). The first approach is typified by using the same measurement tool with the same group of test-takers and still

Original Article / Doi: 10.26750/Vol(10).No(1).Paper11

obtaining relatively similar scores. The second approach is exemplified by questioning the extent to which the scores obtained by using a measurement tool are accurate, and whether the scores are exact scores of test-takers actual ability or not. The third approach is illustrated by investigating how much measurement error is caused by the measurement tool itself, for which there are two types; systematic and random variance (Mousavi, 2012, p.622). In general, unreliability stems from three primary sources; structure of the measurement tool, administration of the testing process and scoring of tests. The first source is due to the improper development of the test items. The second one has a direct relation with variation which might occur due to an inconsistent administration environment. The last one originates from the variation which is caused by a subjective rating of the performance. In the current study the main concern is about the test structure as the other two sources of unreliability (variables) were pretty much controlled; the scoring was done automatically by the software set up for that purpose, the administration environment was the same for all the test takers.

The issue of reliability is best described in Classical Test Theory which posits the notion of test-takers true score; the score which would be obtained by an assessment tool that is a perfectly reliable measure of the candidate’s performance. However, the true score exists only in theory; its approximation is obtainable by gathering samples of performance. The replication of the assessment procedure causes inconsistency of the scores and thus measurement errors. Classical Test Theory reliability coefficients provide an index, as they are widely used in social sciences, ranging from 0 to 1.00 (Webb et al., 2006). The coefficients 0.80 and above are considered to be convenient and crucial for language tests which are devoted to test constructs such as vocabulary and grammar, and reading skill. Yet, Lado (1961) cited in (Hughes, 2003, p.39) said that the reliability coefficient of good vocabulary, structure and reading tests are 0.90 to 0.99. The test structure and sampling, in particular, as the measuring tool, affect the scores to vary significantly. The sampling adequacy is, a quite a long time ago, considered to be one of the factors which affect test reliability for which the split-half method will be suitable to estimate test reliability (Harris, 1969). Similarly, the inter-item consistency can be the area of concern in which test reliability is estimated by the proportion of test-takers pass and failure of each item (Harris, 1969). All the aforementioned sources and factors of measurement errors are widely researched to revise and enhance the measurement tool to make it yield accurate results. This is further motivated by the degree of stakes and the decisions made based on the test results. However, that does not imply if the results of a measurement tool are not very significant should not be reliable. Even in the case of placement tests which are conducted to indicate English language learners’ ability, the stake is high. Bachman and Purpura (2008) maintained that one of the intended uses of language assessment is “to provide score-based information for classifying students … according to their level of language ability so that they can receive level-appropriate instruction.” (p. 458). In other words, what they meant was placing learners into homogenous groups based on ability level or readiness to be engaged in receiving appropriate instruction. Placement tests can also determine whether learners be exempted from attending a specific language level courses (Green, 2012). Thus, decisions are made on the basis of test scores. The researchers also claimed that in order to make fair and equitable decisions, we need to consider the quality of information obtained by the assessment tool; the information should be characterized with reliability and validity (p.456). All things considered, examining and the verification of the test results are invaluable scientific ventures regardless of the degree of significance and stakes foreseen. A close look at the venue of language placement testing research illustrates those issues such as validity and reliability that have received the attention of a considerable number of researchers, and many types of research can be found in the empirical literature. These studies (Aron and Aron, 2003, Bachman, 1990, 2004, 2005, Brown, 2004, Brown, 2005, Hughes, 2003, Piedmont, 2014; among others) proposed not only

Original Article / Doi: 10.26750/Vol(10).No(1).Paper11

one approach but several approaches and methods for assessing and computing the reliability estimates. What follows is an overview of some of these approaches.

2.2 Methods of estimating reliability

Approaches and methods of calculating reliability include reliability coefficient, Rasch Model, Standard Error of Measurement, etc., for which various statistics and formulas are developed (Spearman-Brown Prophecy, Cronbach Alpha and Kuder-Richard 20 formulas). The two most common reliability statistics are the reliability coefficient and the standard error of measurement. They refer to the same sources of inconsistency in the scores obtained by using a measurement tool. First, the reliability coefficient is a measure of the accuracy of a test or measuring instrument obtained by measuring the same individuals twice and computing the correlation of the two sets of measurements (Meriam-Webster Dictionary, n.d.). “The correlation between the sets of observations provides a reliability coefficient” (Piedmont, 2014). It is an index of the amount of true variance operating in a set of raw test scores (Aron and Aron, 2003).

The following is the Reliability Coefficient formula:

$$RC = \left(\frac{N}{N-1}\right) \times \left(\frac{Total\ Variance - Sum\ of\ Variance}{Total\ Variance}\right)$$

Secondly, the Standard Error of Measurement (SEM) is “a statistic that is used to estimate limits within which an individual’s OBTAINED SCORE on a test is likely to diverge from his TRUE SCORE” (Mousavi, 2012, p. 694, emphasis in origin). Reliability scholars (Bachman, 1990, 2004, Brown, 2004, Brown, 2005, Hughes, 2003; among others) agree that calculating the true score is not attainable, thus the estimate of it needs to be calculated. Similarly, nearly all of them agree that reliability is about measurement error, be it systematic or random. SEM formula is as follows:

$$SEM = SD\sqrt{1 - r_t}$$

Where SD is the standard deviation and r_t is the estimated reliability.

As far as data collection method is concerned, in general, there are four methods to assess the reliability of a measuring instrument; test-retest reliability, parallel forms reliability, inter-rater reliability and split-half reliability (Bachman and Palmer, 1996, Fulcher, 2013, Hughes, 2003; among others). The split-half method, as it is the primary method to calculate reliability in the current research, concerned with how much error of a test score has resulted from a poor test construction.

Some empirical studies were reviewed that were directly related to the scope and aim of the current study. The studies focus on various aspects of measurement tool reliability; internal consistency, test qualities such validity and reliability, and the content area of the test. For example, Long et al. (2018) investigated the internal consistency and validity of a new web-based Spanish language test. For this purpose, a 100-item test which was distributed across types of items as sound discrimination, grammar, listening comprehension, reading comprehension, and vocabulary was used to test 2,201 incoming first-year students. Analysis of the results revealed that the test is valid and reliable with regard to its function, content coverage, and the consistency of the placement decisions. The study used Bachman’s (2005) assessment use argument framework to examine the evidence which supports the validity and reliability of the placement test under study. Fan and Jin (2020) studied how placement tests were developed, implemented and used to conform to the best practices of language assessment in China. They used a mixed-method for data

Original Article / Doi: 10.26750/Vol(10).No(1).Paper11

collection. The findings suggest a lack of quality control in placement test practice which raises rightful concerns regarding reliability, validity, and usefulness of the test at higher education foreign language programs. Razavipour and Firoozi (2021) researched the placement tests decisions, uses and policies in 30 language institutions in Iran. The focus of their study was on the content area of the test, and the statistical analysis of the data revealed that grammar and vocabulary, and listening skill were the two main language elements tested; meanwhile, reading, writing and translation were not being primary elements in the test. As far as the decisions made were concerned, they were not valid, reliable indicators of learners' language ability which were being influenced by institutional, gender and age factors.

Compared to other foreign language assessments, placement tests are smaller, less frequent and of lower stakes. Placement tests, with their direct and considerable impact on learners, are widely used nowadays in foreign language programs at the university level. However, it has been found that very few studies, to the best of our knowledge, attended to investigating placement test development and administration, with a particular focus on the reliability of the measurement tool. Though a bulk of research studies addressing the issue is available, none of them attended to test reliability evaluation area in Kurdistan Region of Iraq, especially at an institutional level. Thus, the research questions guiding the current study are the following:

1. What is the reliability estimate of the University of Raparin Language and Development Center Placement Test?
2. To what extent is the University of Raparin Language and Development Center Placement Test a reliable indicator of the English language level of the incoming first-year students?
3. To what extent are inferences drawn from the University of Raparin Language and Development Center Placement Test scores meaningful?

This type of research is essential as it attempts to provide pieces of evidence that supports the reliability and validity of the inferences and decisions made based on the scores obtained by administering the UoRLDC placement test.

3. Methodology

The method used in the current study to compute the reliability coefficient of the test is the split-halves method (Hughes, 2003). In this method, unlike the test-retest and alternative-form methods, the reliability is calculated by conducting the test on one occasion only. The technique is to divide the entire set of similar items into halves, and the scores on the halves are correlated to yield an estimate of reliability (Carmines and Woods, 2005). There are many statistical methods to calculate the reliability estimates; the correlation between the two sets of scores (Fulcher, 2013). The most common and most straightforward strategy to use is the split-half method, in which two sets of scores can be obtained from one single administration of a test. The use of the method offers a type of coefficient which belongs to the internal consistency (Hughes, 2003), which is the primary goal for the current research. Here, the 40 items of the UoRLDC Placement Test are divided into two halves based on odd and even-numbered items. Each of the halves is considered an estimate of alternative forms. However, the correlation between the reliability of the two halves is the reliability for each half but not for the entire test. Carmines and Woods (2005) proposed a statistical correction, which is called the Spearman–Brown prophecy formula, to estimate the reliability of the entire test. They also suggested that researchers, using the Spearman-Brown prophecy formula, can determine the number of items needed to attain a given reliability rate. The formula and its implementation are illustrated in the next section.

Original Article / Doi: 10.26750/Vol(10).No(1).Paper11

Additionally, the method in the research is to deal with two types of statistical information; the first relates to the tests as a whole and the second to the individual sections which constitute the test. Another statistical analysis is done by using Kuder-Richard 20 formulas to compute the reliability estimate of the test if its items are increased.

Furthermore, frequency of the scores, mean, standard deviation, standard error of measurement and confidence intervals are also calculated to provide the needed data for internal-consistency reliability estimate. The goal is to compare the results of reliability estimates obtained by utilizing each of the aforementioned formulas and draw conclusions in the light of their results.

4.1 Data Analysis and Discussion of the Obtained Results

In this section, the data from the UoRLDC Placement Test is analyzed and graphically displayed. First, the analysis starts with showing the frequency distribution of the results, which was between 17 to 19 grades out of 40. Additionally, the descriptive data such as the mean, STD, STD error computed by Microsoft Excel Sheet are also shown in the Figure and Table 4.1. below:

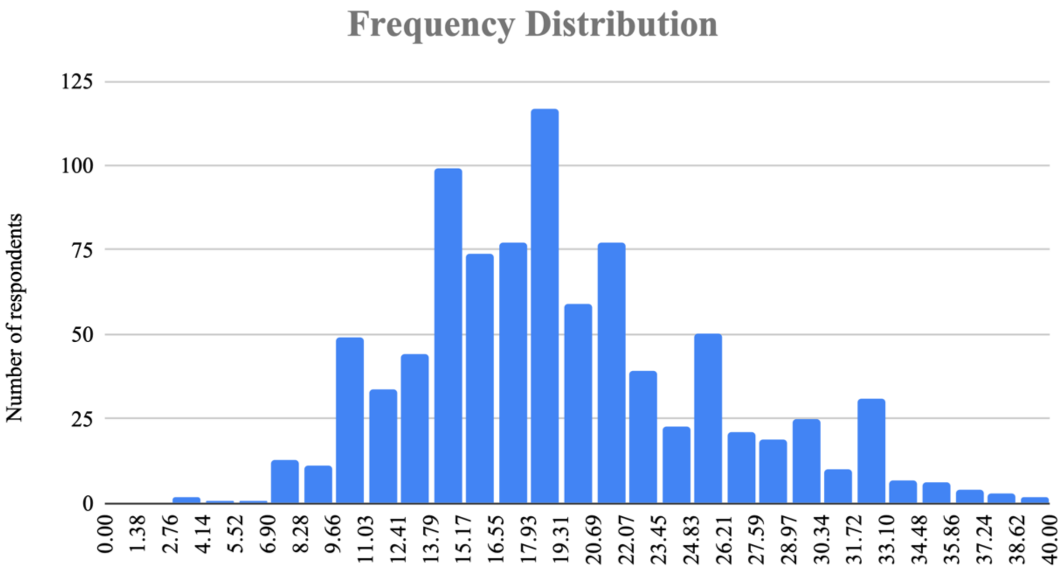

Figure 1. Frequency Distribution of Grades out of 40

Table 4.1. Illustration of the descriptive data for the entire UoRLDC Placement Test Scores.

Descriptive Data	Numbers
Mean	19.27171492
STD	6.3570554
STD ERROR	0.2121376863

Secondly, the reliability coefficient, the mean and variance of each of the two halves (T1, T2) are calculated by Microsoft Excel Sheet[ii], as shown in Table 4.2 below. The results show the UoRLDC Placement Test confidence interval is 63%, and the range of true score to the observed one is two grades. With regard to the odd/even split of items results in two identical halves in terms of content and other item characteristics, it is obvious the items in T1 and T2 are similar. The number of items in each half is 20 items, the testing technique used is multiple-choice questions, and the means vary slightly. However, given the single item difficulty, it is not the aim of the current study to calculate it.

Table 4.2. Illustration of URLDC Placement Test Exam Scores of T1 and T2 tests.

Descriptive Data	T1 (Odd-Numbered)	T2 (Even-Numbered)
Total Scores	9694	7612
The mean	10.8	8.5
Variance	12.54659804	11.51617493
Correlation between T1 & T2	0.6800708875	

The reliability estimate for each of the halves is 0.68. However, this is not the reliability of the entire test. The Spearman-Brown prophecy formula is employed to calculate the reliability of the whole test. As the number of the whole test is two times as long as the halves, Carmines and Woods (2005) maintain that the formula should be expressed as the following:

$$\rho xx'' = \left(\frac{2\rho xx'}{1+\rho xx'} \right)$$

where $\rho xx''$ is the reliability coefficient of the entire test, while $\rho xx'$ is the correlation between the two halves and the reliability estimate for each. Thus, the reliability estimate of the whole test is calculated like the following:

$$\rho xx'' = \left(\frac{(2)(0.68)}{(1+0.68)} \right) = \left(\frac{1.36}{1.68} \right) = 0.81$$

Thirdly, the reliability of the test scores is also calculated by Kuder-Richardson (KR20) formula shown below to verify the scores reliability estimate:

$$r_{KR20} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma pq}{\sigma^2}\right)$$

Where K is the number of the items (40), σ^2 is the variation of the total exam scores (40.41), and Σpq is the proportions of both the pass and failed test-takers (8.235). Each value had been calculated separately. For example, the variation of the total exam scores was calculated by squaring the standard deviation of the scores, i.e., the dispersion of the scores from the mean. Similarly, the proportions of both the pass and failed test-takers were summed up to obtain the value required for the Kuder-Richardson formula[iii].

The reliability estimate of the placement test is 0.81. As it is shown below, the result obtained by the Kuder-Richardson formula is calculated as:

$$r_{KR20} = \left(\frac{40}{40-1}\right)\left(1 - \frac{8.235}{40.41}\right)$$

Original Article / Doi: 10.26750/Vol(10).No(1).Paper11

$$r_{KR20} = 0.81$$

Nonetheless, the reliability estimates for each of the component parts (grammar & vocabulary and reading) of the test by implementing the formula are also calculated:

$$r_{KR20} = (\frac{34}{34-1})(1 - \frac{7.242}{34.574})$$

$$r_{KR20} = 0.814318 \textit{ (grammar \& vocabulary)}$$

$$r_{KR20} = (\frac{6}{6-1})(1 - \frac{0.99}{1.44})$$

$$r_{KR20} = \left(\frac{6}{5}\right)(1 - 0.6875)$$

$$r_{RK20} = (1.2)(0.3125) = 0.37 \textit{ (Reading)}$$

The results show that the test scores reliability did not reach the desired conventional reliability supported by research (Lado 1963 as an example). The results also demonstrated that the test structure, particularly the number of items in the test (the reading part of the test which has six items), was the primary variable which caused that low level of reliability. For instance, the number of items needed to raise the reliability level of the overall test to above 0.90 was pretty low; they should be increased to 92 items instead of 40. The Spearman–Brown prophecy formula proposed a formula to determine the number of items that would be needed to attain a given desired reliability. "To estimate the number of items required to obtain a particular reliability, the following formula is used" (Carmines and Woods, 2005, pp. 364):

$$N = \frac{\rho xx''^{(1-\rho xx')}}{\rho xx'^{(1-\rho xx'')}}$$

$$N = \frac{0.90(1-0.81)}{0.81(1-0.90)} = \frac{0.171}{0.081} = 2.1$$

> where $\rho_{xx''}$ is the desired reliability, $\rho_{xx'}$ is the reliability of the existing test, and N is the number of times the test would be lengthened to obtain the reliability of $\rho_{xx''}$. For example, if a 10-item test has a reliability of 0.60, then the estimated lengthening required to obtain a reliability of 0.80 would be $N = 0.8(1 - 0.6)/0.6(1 - 0.8) = 2.7$. In other words, 27 similar items are required to attain a reliability rate of 0.80. (Carmines and Woods, 2005, pp. 364).

Hence, the desired reliability for the reading part of the test, if taken separately, based on the results of the empirical studies carried out in the field (Lado, 1961, for example), needs to be no less than 0.90. However, the current reliability estimate is $r_{RK20} = 0.37$, and it implies that in order to attain the desired reliability estimate, which can also be calculated by the Spearman-Brown prophecy formula, the number of the items needed should be no less than 92 identical items. This is shown in the below calculation; the estimated number of items in the test required to obtain 0.90 would be:

$$N = 0.9(1 - 0.37)/0.37(1 - 0.9) = 15.32;\ 15.32*6= 91.94$$

Discussing the obtained results, the test reliability estimate is 0.81. It is convenient as the decisions made based on the results were not highly significant – they are of low or medium-stakes nature. In other words, the results support, to a great extent, the inferences drawn from them, which much serve in achieving the goals for which the test was designed. It was designed to place the incoming first-year students into the language classes which best suit their language ability. Yet, as a criterion-referenced test, it was designed to assess the test-takers' vocabulary and grammar knowledge and reading skill, but not their actual language ability. Hence, the scores might not be the actual indicators of those learners' language ability. The reason is that the content area of the test, as it does not include listening, speaking and writing parts, lacks characteristics of a full-featured placement test by which significant decisions can be made.

Additionally, based on the results obtained, it can be inferred that the test-takers observed scores obtained from the test would be (80%) similar to the ones they will get if they take an alternative form of the test. This is based on the claim the split-halves method division of the total test items into two halves can relatively be considered the alternative forms of the test items. Thus, the reliability estimate of the tests' constructs is further supported and verified by the fact that if an alternative form of the test was used, it would still have a convenient reliability estimate. Therefore, the inferences drawn from the scores, though they were not compared against specified benchmarks, are meaningful. However, as the benchmark towards which the university's English language policy is directed is B1, according to Common European Framework of Reference, the UoRLDC Placement Test does not have the power to exempt test-takers who got high grades from taking the courses. Then, the inferences drawn are not fully meaningful by which decisions can eventually be made.

Furthermore, based on the results, the current structure of the UoRLDC Placement Test (test structure factor) needs to be modified, with particular focus on the number of test items. Increasing the items will characterize the test to be qualified as the real indicator of learners' language level. Consequently, decisions such as learners with high scores exemption from taking the language courses can be taken. This can be generalized to other language measurement tools, since the obtained scores by these tools are used to draw inferences, and consequent decisions about test-takers are taken. Hence, examining and checking on the measurement tools should be made a convention to get results that are fair and equitable on the one hand, and reflect the test-takers real performance on the other.

4.2 Conclusion

In general, the UoRLDC Placement Test has a convenient reliability estimate of 0.81. With the current form, the test can achieve the goals for which it was initially designed. That is to say, it is roughly a reliable indicator of learners' vocabulary and grammar knowledge as well as their reading skill. However, if taken separately, the reliability estimates for the single constructs of grammar and vocabulary, and reading skill were not high. The test is not a reliable indicator of learners' reading skill. Test construct variable, as one of the factors affecting the reliability of test scores, contributed to the low-reliability estimate, notably, the number of the items for each of the constructs in the UoRLDC Placement Test. Hence, the test needs to be modified and improved to yield scores by which meaningful inferences are drawn, and significant decisions can be made thereafter. For example, the vocabulary and grammar test section needs a greater number of items (the number should be doubled) to obtain the desired reliability estimate. Nonetheless, the number of items devoted to assessing learners' reading ability should also be increased by nearly five times due to the very low-reliability estimate. It is recommended that the number of the items in the test should not be less than 100 items to obtain the desired reliability estimate of above 0.90, as it was proposed and supported by empirical studies long before.

The case attended in the study can help language teachers, educators, test administrators, etc. begin thinking about changing the current convention of getting a sample of learners' performance with a tool that may have specific problems in terms of its construction, administration and scoring. The measurement tools themselves need to be liable to constant examination and check before or after they are put into a widespread use. Last but not least, the study investigated a variable among the countless number of factors that influence the quality of the test to yield reliable results as a case study. Other cases and aspects, such as test administration and scoring, also need to be investigated constantly. Meanwhile, the validity of human cognition measurement tools should also be studied as a significant feature for a test to gain accurate results which can reflect the candidates' language ability and performance. All the procedures of scrutinizing and checking carried out on measurement tools have a single aim; it is to make the assessment fair and equitable. Hence, the study results would suggest further investigation into all the placement tests at all the universities in Kurdistan Region to check their validity and reliability.

لێکۆڵینەوە لە کواڵێتی تاقیکردنەوەکانی زمان: بژاردنی ژمارەی جێگیری تاقیکردنەوەی دیاریکردنی ئاست وەک نمونە

جمال علی عمر[١] - سەردار عبدالله حسێن[٢]

[١+٢]بەشی زمانی ئینگلیزی، کۆلێژی پەروەردەی بنەڕەت، زانکۆی ڕاپەڕین، ڕانیە، هەرێمی کوردستان، عێراق.

پوختە:

ئامڕازەکانی پێوانەکردنی توانستی زمان دەبێ متمانەپێکراو و ئەنجامەکانیان جێگیر و پشت پێ بەستراو بن، بەو پێیەی کە ئەو نمرانەی لەڕێگای ئەوانەوە بەدەستهاتوون هاوکار دەبن لە گەیشتن بە دەرئەنجام و بڕیاردان لەسەر بەشدار بوانی تاقیکردنەوەکان. ئەم توێژینەوەیە لە جێگیری و پشت بەستراوی ئەنجامی تاقیکردنەوەی دیاریکردنی ئاستی زمانی ئینگلیزی دەکۆڵێتەوە. ئامانجی توێژینەوەکە بریتیە لە دۆزینەوەی ڕێژەی جێگیری ئەنجامەکانی ئەو جۆرە تاقیکردنەوانە. لەسەر بنەمای تیۆری تاقیکردنەوەی کلاسیکی، کارتێکەری جێگیری (پشت پێ بەستراوی) تاقیکردنەوە پێویستە شیکار بکرێ بۆ دۆزینەوەی مەودای هەڵەکردن لە پێوانەدا و بەدەستهێنانی خەملاندنی نمرەی ڕاستەقینەی بەشداربوێک. بۆ ئەم مەبەستە، ئەو داتا و زانیاریانە شیکراونەوە کە لە ئەنجامدانی تاقیکردنەوەی دیاری کردنی ئاست کە لە سەنتەری زمانی زانکۆی ڕاپەڕین بۆ ٨٨٩ خوێندکاری قۆناغی یەکەم ئەنجام درا. توێژینەوەکە میتۆدی شیکاری دیارخەری بەکارهێنا کە تێیدا بە بەکارهێنانی شێوازی دابەشکراو بۆ نیوە هاوکێشەکانی کۆدەر–ڕیچارد سۆن ٢٠ و سپێرمان–براون پەیام هێنەر بەکارهێنراون بۆ دۆزینەوەی ڕێژەی خەمڵێنراوی جێگیری تاقیکردنەوەکە. ئەنجامەکان دەریان خست کە ڕێژەی جێگیری خەمڵێنراوی تاقیکردنەوەکە ٠,٨ ە. هەروەها ئەنجامەکان نیشانیاندا کە بۆ بەدەستهێنانی ڕێژەی جێگیری ٠,٩ یان سەرو پێویستە ژمارەی بڕگەکانی تاقیکردنەوەکە بکرێن بە دوو ئەوەندە. بەمشیوەیە، بە پشتبەستن بە ئەنجامەکانی توێژینەوەکە، پێشنیار دەکرێ کە پێکهاتەی تاقیکردنەوەکە باشتر بکرێ بۆ بەدەستهێنانی ڕێژەی خەمڵێنراوی جێگیری بەرز تر. وەهەروەها پێشنیار دەکرێ کە تاقیکردنەوەکە دەبێ دەستکاری بکری؛ بڕگەکانی زیاد بکرێن، ناوەڕۆکەکەی چاکبکرێت بۆ ئەوەی هەموو کارامەییەکانی بەکارهێنانی زمان هەڵبسەنگێنی. لەوەش زیاتر، پێشنیار دەکرێ کە لیکۆڵینەوە لە سەرچاوەکانی تری ناجێگیری ئەنجامەکان لە نێوانیاندا بەڕێوەبردنی تاقیکردنەوەکان و تاک تاکی پرسیارەکانی تاقیکردنەوەکە ئەنجام بدرێ بۆ دۆزینەوەی سەرچاوەی هەڵە لە پێوانەکردندا.

کلیلە وشەکان: کارتێکەری جێگیری (پشت پێ بەستراوی) ئەنجامەکان، شێوازی دابەشکراو بۆ نیوە، تاقیکردنەوەی دیاری کردنی ئاست.

Original Article / Doi: 10.26750/Vol(10).No(1).Paper11

References

ARON, A. & ARON, E. N. 2003 (3rd ed.) *Statistics for psychology*. United States of AMerica: Prentice-Hall/Pearson Education.

ARYADOUST, V. & RAQUEL, M. 2019. *Quantitative data analysis for language assessment Volume I: Fundamental techniques*. USA: Routledge; Taylor & Francis Group.

BACHMAN, L. F. 1990. *Fundamental considerations in language testing,* Oxford, England, Oxford University Press.

BACHMAN, L. F. 2004. *Statistical analyses for language assessment*, Oxford, England, Oxford University Press.

BACHMAN, L. F. 2005. Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal,* 2**,** 1-34.

BACHMAN, L. F. & PALMER, A. S. 1996. *Language testing in practice: Designing and developing useful language tests*, Oxford University Press.

BACHMAN, L. F. & PURPURA, J. E. 2008. 32 Language Assessments: Gate-Keepers or Door-Openers? *The handbook of educational linguistics***,** 456-469.

BROWN, H. D. 2004. Language Assessment: Principles and Classroom Practices: United Stated of America. A Pearson Education. Inc.

BROWN, J. D. 2005. *Testing in language programs: a comprehensive guide to English language assessement,* New York, McGraw-Hill College.

CARMINES, E. G. & WOODS, J. A. 2005. Reliability Assessment. *In:* KEMPF-LEONARD, K. (ed.) *Encyclopedia of Social Measurement.* New York: Elsevier.

CHAPELLE, C. 2013. Reliability in language assessment. *In:* CHAPELLE, C. A. (ed.) *The Encyclopedia of Applied Linguistics.* Oxford: Blackwell/Wiley.

FAN, J. & JIN, Y. 2020. Standards for language assessment: demystifying university-level English placement testing in China. *Asia Pacific Journal of Education,* 40**,** 386-400.

FULCHER, G. 2013. *Practical language testing*, Routledge.

GREEN, A. 2012. Placement testing. *In:* C. COOMBE, P. D., & B. O'SULLIVAN (ed.) *The Cambridge guide to second language assessment.* New York: Cambridge University Press.

HARRIS, D. P. 1969. Testing English as a Second Language.

HUGHES, A. 2003. *Testing for language teachers (2nd Edition)*, Ernst Klett Sprachen.

KUDER, G. F. & RICHARDSON, M. W. 1937. The theory of the estimation of test reliability. *Psychometrika,* 2**,** 151-160.

LADO, R. 1961. Language Testing: The Construction and Use of Foreign Language Tests. A Teacher's Book.

LONG, A. Y., SHIN, S.-Y., GEESLIN, K. & WILLIS, E. W. 2018. Does the test work? Evaluating a web-based language placement test. *Language Learning & Technology,* 22**,** 137-156.

MERIAM-WEBSTER DICTIONARY. n.d. *Reliability coefficient* [Online]. Available: https://www.merriam-webster.com/dictionary/reliability%20coefficient [Accessed February 4 2021].

MOUSAVI, S. A. 2012. *An encyclopaedic dictionary of language testing,* Tehran: Iran, RAHNAMA PRESS.

PIEDMONT, R. L. 2014. Reliability Coefficient. *In:* MICHALOS, A. C. (ed.) *Encyclopedia of quality of life and well-being research.* Springer Netherlands Dordrecht.

RAHMAN, M. S. 2020. The advantages and disadvantages of using qualitative and quantitative approaches and methods in language "testing and assessment" research: A literature review.

RAZAVIPOUR, K. & FIROOZI, T. 2021. Placement Decisions in Private Language Schools in Iran. *Challenges in Language Testing Around the World: Insights for language test users***,** 267.

WEBB, N. M., SHAVELSON, R. J. & HAERTEL, E. H. 2006. Reliability Coefficients and Generalizability Theory. *Handbook of Statistics*, 26.

Endnotes:

[i] https://docs.google.com/spreadsheets/d/150wLN23qjQy4Dyp19MeA0qcRDtRYkXJHRNEJRGLhE_s/edit#gid=232600345

[ii] https://docs.google.com/spreadsheets/d/150wLN23qjQy4Dyp19MeA0qcRDtRYkXJHRNEJRGLhE_s/edit#gid=1564386594

[iii] https://docs.google.com/spreadsheets/d/150wLN23qjQy4Dyp19MeA0qcRDtRYkXJHRNEJRGLhE_s/edit#gid=1564386594