International Journal of Linguistics and Translation Studies
Volume 5, Issue 2, 2024
Homepage: http://ijlts.org/index.php/ijlts/index
DOI: https://doi.org/10.36892/ijlts.v5i2.430

Evaluating Test-Retest Reliability of Language Tests in Moroccan Tertiary Education

Anouar Mohamed El Kasri [1]
Faculty of Arts and Humanities, Moulay Ismail University, Meknes, Morocco
anouar.elkasri@yahoo.com

Mohammed Larouz
Faculty of Arts and Humanities, Moulay Ismail University, Meknes, Morocco

Moulay Sadik Maliki
Faculty of Arts and Humanities, Ain Chock, Hassan II University, Casablanca, Morocco

Brahim El Yousfi
National Institute of Agricultural Research (INRA), Morocco

How to cite: El Kasri, A. M., Larouz, M., Maliki, M. S., & El Yousfi, B. (2024). EVALUATING TEST-RETEST RELIABILITY OF LANGUAGE TESTS IN MOROCCAN TERTIATY EDUCATION. *International Journal of Linguistics and Translation Studies*, *5*(2), 41–53. https://doi.org/10.36892/ijlts.v5i2.430

ARTICLE HISTORY

Received: 23/01/2024

Accepted: 27/02/2024

KEYWORDS

reliability, test-retest method, variance, language tests, Morocco

Abstract

Reliability is an essential element of assessment principles, and various methods were developed to measure language tests' reliability, including test-retest. This method is widely discussed in the literature; however, no studies investigating test-retest reliability have been published in Morocco. Therefore, the present study evaluated the test-retest method in the Departments of English Studies in the Moroccan Faculties of Letters and Humanities. One thousand seven hundred and seventy-two semester one students from three Moroccan universities took the grammar, paragraph writing, spoken English, and study skills tests on two occasions. The data consisting of students' scores in these modules were analyzed using descriptive statistics and reliability test analysis. The findings showed that variance among students and between test and retest scores was highly significant (p=0.001), and revealed that faculty, modules, and their interaction significantly affected the difference between test and retest scores. The implications of these results and the perspectives are discussed in the text.

1. INTRODUCTION

Assessment is an important component of the teaching and learning process, and ensuring its quality is a crucial issue in education. Assessment is the only way to determine whether teaching activities have resulted in effective learning outcomes. However, many teachers and test designers find it challenging to construct useful tests (Gronlund, 1982). In this regard, a

[1] Corresponding author: anouar.elkasri@yahoo.com

study showed that nine out of twelve U.K. tests reviewed did not provide sufficient evidence of reliability and validity (Alderson & Hamp-Lyons, 1996). In Morocco, institutional and instructional practices revealed that teachers needed to understand assessment and testing procedures due to the absence of any framework explaining how they should be implemented (Ghaicha, 2016). Accordingly, comprehensive test analysis in Moroccan faculties showed a higher need for levels of reliability and validity (Abouabdelkader, 2018).

In this regard, many frameworks have been developed to design tests that reflect language assessment principles. Brown (2004) developed a model for 'Testing a test' with five cardinal language assessment criteria that test designers should respect while creating language tests. These criteria are reliability, authenticity, practicality, validity, and washback. Brown's model was based on another model called 'Test usefulness', which had already been developed in 1996 by Bachman and Palmer. Besides the first three criteria, this model includes construct validity and impact instead of validity and washback and a new criterion called interactiveness. The reliability criterion remains paramount in designing practical language tests in both 'Test usefulness' and 'Testing a test' models.

The major goal of this paper is to measure the reliability of language tests for Moroccan common-core students at three Moroccan Faculties of Letters and Humanities. In this context, we used a test-retest method to determine how reliable the language tests designed for those students in these faculties are. This study comprises three main sections. Section one offers a theoretical background delineating the different types of reliability and the challenges related to their application. The procedural techniques adopted in this study along with a description of the tests exploited to investigate the issue under prime consideration are outlined in section two. The third section presents the results of this research and a discussion of its findings. The study ends with some implications for curriculum designers and test developers as well as research perspectives and recommendations for future research.

2. LITERATURE REVIEW

Reliability refers to the consistency of test scores (Harris, 1969; Gronlund, 1977; Coombe et al., 2007; Brigui, 2017). Regardless of the similarity between them, two tests will likely consistently produce different results (Brigui, 2017). The reliability of test scores depends on how comparable these scores are (Wells & Wollack, 2003). Furthermore, a test is reliable when it generates the same outcomes on different occasions under the same exam conditions (Richards, 2015). In other words, test reliability depends on the ability of good language tests to yield consistent results; tests should measure consistently whatever they are supposed to measure under all conditions (Maduekwe, 2007). To ensure assessment task reliability, which is one of the prominent language assessment criteria, different methods were discussed in the literature such as parallel, internal consistency, inter/intra-rater, and test-retest reliability (Gipps, 1994; Strong, 1995; Brown, 1996; Chiedu & Omenogor, 2014; Neukrug & Fawcett, 2015).

2.1. Parallel or Alternate Form Reliability

One common means to determine the reliability of language tests is to develop two or more parallel, alternate, or equivalent forms of the same test. For two tests to be considered parallel, Bachman (1990) assumes that they measure the same ability, which means that the student's true scores on one test will be similar to the other test scores. This method uses two different but equivalent forms of assessment, which are administered to the same group during the same

testing session. The two forms are also built to the same specifications but constructed independently (Gronlund, 1977; McMillan, 2018). This assumes according to Bachman and Palmer (1996) that although two similar versions of a particular test can be constructed, they cannot be regarded as parallel unless statistical analysis of the tests and all the items are included. The correlation between the scores on the two forms of the test is considered a measure of reliability (Fulcher & Davidson, 2007). The challenge in this type of reliability is to ensure that the forms of the same test use the same or very similar directions, format, and number of questions, and are equal in difficulty and content. (Neukrug& Fawcett, 2015).

2.2. Internal consistency

To avoid the complexity of assuring reliability associated with parallel forms strategies, testers usually resort to another type of reliability called internal consistency or internal consistency reliability evidence. Unlike the previous form, this type of reliability deals with the extent to which the items in a test function consistently rather than a mere focus on the consistency of students' scores (Popham, 2017). This type of reliability is called internal consistency because we are not going outside of the test; instead, we look within the test to determine reliability estimates (Neukrug& Fawcett, 2015). In other words, internal consistency requires only one form of a test and a unique administration of this form to ensure the estimation of its reliability (Brown, 1996). Among the types of internal consistency reliability, the most basic form referred to in the literature is called split-half or odd-even reliability.

The split-half method is based on the principle that if a measuring instrument is divided into two halves, the measurements obtained in both parts would correspond to each other (Heaton, 1990). Accordingly, this method consists of dividing the test into two halves and then determining the extent to which the scores of these two halves are consistent. In so doing, the two halves are treated as parallel tests (Brown, 1996). Students take the test, and each one is given two scores; one score on each half. These scores are then used to get the reliability coefficient as if the whole test had been taken twice (Hughes, 1989).

2.3.Rater reliability

Two other types of reliability are necessary in language testing, where raters judge students' responses. During this scoring process, human error, subjectivity, and biased judgment may take place. According to Richards (2015), if two raters use a checklist to assess a student's essay, for instance, and give completely different grades, the checklist would lack reliability. It would also lack reliability if a teacher gave one set of marks using the checklist on one occasion and then assessed the same piece of writing on another occasion and gave it a different grade. Richards is referring here to two types of rater reliability extensively discussed in the literature, which are inter-rater and intra-rater reliability.

Inter-rater reliability is a variation of the equivalent-forms type of reliability since at least two raters produce the scores before calculating a correlation coefficient between them (Brown, 1990). This method of reliability aims to evaluate the degree to which different test items that probe the same construct produce similar results (Chiedu & Omenogor, 2014). In other words, this type of reliability requires that two or more raters of the same test give consistent scores (Brown, 2004). Because raters have different circumstances that are likely to affect their judgment, including their physical conditions, emotional state, etc., different raters will not necessarily interpret students’ answers in the same way. Besides, measures that use objectivity

in scoring, such as true-false questions, multiple-choice, etc., tend to produce higher agreement among scorers than those relying on more subjective scoring methods like essays (Erford & Bardhoshi, 2017).

On the other hand, intra-rater reliability is a procedure that is not limited to situations where two or more scorers are involved in evaluating students' answers. It is another reliability measure that occurs when one rater scores consistently from one student's paper to another. Intra-rater reliability is similar to the test-retest strategy in the sense that the same scorer produces two sets of scores, for the same group of students, on two different occasions, before a correlation coefficient is calculated (Brown, 1996). Brown (2004) contends that unclear scoring criteria, fatigue, bias toward particular good and bad students, or carelessness may affect intra-rater reliability. Nevertheless, if scorers' judgments of a language performance sample, whether written or spoken, are based on a set of criteria of what constitutes an adequate performance, this will yield a reliable rating set (Bachman, 1990). Rater reliability in writing tests is particularly hard to achieve since writing proficiency involves numerous traits that are difficult to define. However, carefully specifying an analytical scoring instrument can increase both intra and inter-rater reliability (Barkaoui, 2011).

2.4. Test-retest reliability

The standard way to assess the reliability of a test is to administer this test a few days apart (Gipps, 1994). Test-retest reliability operates through three basic steps. First, the test aims to measure the test reliability for the same group of students, twice and under the same conditions (Heaton, 1990; Bachman, 1990; Brown, 1996; Chiedu & Omenogor, 2014). The period between these times differs according to the results' interpretations (Gronlund, 1977). Next, the correlation coefficient of the two scores is calculated and interpreted. When test takers are not allowed to review or practice the target content, the test is likely to produce different results and become unreliable (Bachman & Palmer, 1996).

Many factors that affect the reliability of language tests have been discussed in the literature. Coombe et al., (2007) contend that three crucial factors affect test reliability. These are the formats and content of the questions and the time given for students to take the exam. In this regard, testing research confirms that longer exams produce more reliable results than brief quizzes (Bachman, 1990). Similarly, Sattler (2001) states that the length is a significant factor in the reliability of tests. That is to say, the longer the test is, the more reliable it becomes. Other factors influencing the reliability are the clarity of the test instructions, whether the objective scoring of the test is possible, the familiarity of the scorers with the test, and the circumstances in which the test is administered (Hughes, 2003). Coombe et al., (2007) also stress the administrative factors claiming that these include the classroom setting (lighting, seating arrangements, acoustics, etc.), and how the teacher manages the administration of the exam. Reliability can also be problematic when a test is a speed test because not every student can complete all the items in this type of test. In contrast, a power test in which every student can complete all the items should be used (Chiedu & Omenogor, 2014).

Test-retest reliability, it should be noted, has some drawbacks; when the time between the first and second administrations is short, students may recall items and their responses, making the same responses more likely and the reliability spuriously high (Hughes, 1989; Alderson & Banerjee, 2002). At the same time, if the period is long, bigger changes could occur affecting

the measured construct (Simkin & Kuechler, 2005). Memory and fatigue effects can also lower students' performance in the second period. Either students may feel exhausted, or their memory may not allow them to recall their answers in the first period. Motivation and maturity effects are also crucial factors causing changes in the test takers' responses or performance over time, which can reduce test-retest reliability. Therefore, to ensure the reliability of language tests, the effect of these factors must be kept at a minimum.

Conducting a study on the test-retest method is a challenging task, especially in the Moroccan context, for many reasons. First, most teachers have to teach an extended program in a few sessions, so it is not easy to convince them to devote two sessions to research. Second, while they find the first test administration expected and logical, students deem it meaningless to sit for the same test another time. Third, at the end of the semester, it is difficult to find enough students to administer the tests. Finally, a common problem most researchers encounter in Morocco is gathering data from educational institutions. Because of that, it is urgent to reconsider the role of researchers and scientific research in reforming our educational system.

To our knowledge, apart from high-stakes tests such as Duolingo and TOEFL, more studies should be conducted to measure language test reliability using the test-retest method. Moreover, no studies measuring test-retest reliability have been published for the Moroccan tertiary education context. Therefore, the present study aimed to measure language test reliability for Moroccan common core students at three Moroccan Faculties of Letters and Humanities. In this context, we used a test-retest method to determine how reliable the language tests designed for common core students at Moroccan Faculties of Letters and Humanities are. The test-retest method in this study aims to measure the reliability of language tests in three Moroccan Faculties of Letters and Humanities.

3. MATERIALS AND METHODS

3.1.Participants

The participants in the present study were Moroccan students in the first semester from the Departments of English Studies at three Faculties of Letters and Humanities: Moulay Ismail University, Meknes; Hassan II University, Ain Chock, Casablanca, and Ibn Tofail University, Kenitra. One thousand seven hundred and seventy-two students from these faculties took the grammar, paragraph writing, spoken English, and study skills tests on two occasions (test and retest) with an interval of ten days. The same teachers conducted and corrected the test and retest for each module to ensure high reliability and low bias.

3.2.Description of the tests

The tests investigated in the present study were grammar, paragraph writing, spoken English, and study skills. This section briefly describes each test in the three faculties. The grammar tests designed at the Moulay Ismail and Hassan II faculties were similar. Students answered four fill-in-the-gap questions about tenses, articles, conditionals, and pronouns. However, the test given to students at Ibn Tofail Faculty contained forty multiple-choice questions about articles, prepositions, quantifiers, compound adjectives, and parts of speech.

The paragraph writing tests were different among faculties. The test at Moulay Ismail faculty contained four tasks. In the first one, students had to write four sentences, using appropriate punctuation and capitalization. In the second and third tasks, they were asked to provide the

topic sentence, the concluding sentence, and three supporting sentences to develop the topic. Finally, the last task asked students to write a paragraph on only one of the suggested topics.

The paragraph-writing test at Ain Chock faculty was different and included three tasks. In the first one, students rewrote five sentences using appropriate punctuation and capitalization. In the second, they had to rewrite the suggested sentences using the words between brackets (not only … but also, in addition, regardless of, despite). In the last one, students reformulated a paragraph on one of the suggested three topics.

The paragraph-writing test designed at Ibn Tofail faculty consisted of twenty multiple-choice questions. The responses to these questions identified error type, sentence type, and incorrect sentence in a group of sentences. Some questions also sought to choose the best topic sentence for the suggested paragraphs, identify the irrelevant sentence in a paragraph, mark the correct option to improve a paragraph, and finally identify a sentence.

The spoken English tests were also different among faculties. The Moulay Ismail faculty test included three tasks. The first one required students to provide the phonetic transcription of the words suggested (e.g., single, check, cure, laugh, etc.). In the second, students rewrote the sentences in orthographic transcription. Finally, students defined the idiomatic expressions in the suggested text.

At Ain Chock faculty, the spoken English test also had three main tasks. First, students had to transcribe the words in the three suggested lists, identify the odd one in each list, and provide the minimal pair to the odd words. In the second question, students provided the IPA transcription of the words satisfactory, independent, critical, and revelation, showed the syllable boundaries, and assigned primary stress to each word. Finally, students had to provide the IPA transcription of the suggested passage and transcribe the underlined passages.

The spoken English test at Ibn Tofail faculty contained forty multiple-choice questions related to the identification of words that contained diphthongs and silent letters. Students also had to mark the correct transcription and syllabification of the words. Then, they were asked to identify words containing long vowels, spot the odd phoneme, recognize 'ed' endings transcribed similarly, identify words with a different stress pattern, and identify words pronounced differently.

The last tests for this study were study skills tests that were differently developed. At Moulay Ismail faculty, students responded to three tasks. In the first one, they had to read eight statements and comment by True (T) or False (F). The second task asked students to circle the best answer to the suggested questions. The last task contained direct questions that students were asked to answer (e.g., 1- what is the difference between intrinsic and extrinsic motivation? 2- Using your words, explain the difference between a dream and a goal.).

At Ain Chock and Ibn Tofail faculties, the methods adopted were different. At Ain Chock, the test included two tasks. In the first one, students had to answer ten direct questions about note-taking, time management, motivation, stress management, and the importance of reading. In the second, they had to paraphrase the suggested passage. At Ibn Tofail faculty, the study skills test comprised thirty multiple-choice questions about time management, test-taking and

studying strategies, effective ways to manage stress, reading sub-skills, strategies to motivate students, and the best ways to be successful as a university student.

3.3. Data collection

The tests used in the present study were collected from the three faculties in the fall semester of the academic year (2021/2022). One year later, professors teaching English modules within the three faculties gave their students the tests a few days before the official exams. Students were invited for the first test at the end of the term and ten days later, they retook the same test (retest). The tests were administered and corrected by the same teachers on both occasions. For reliability analysis of the test and retest scores, we applied the reliability procedure using test and retest variables as items and the statistic submenu to evaluate descriptive statistics, interclass correlation coefficient and a two-way random model with absolute agreement as type. We also measured the difference between the tests and retest scores (retest–test scores) for each module by faculty. Moreover, to depict the interaction effect, differences were multiplied by -1. Afterwards, the transformed differences were tested for the effect of modules, faculty, and their interaction significance using mixed model analysis. This later model was herein equivalent to GLM (because the two effects and their interaction were taken as fixed variables). Note that all data analyses were performed by the SPSS software version 26 (IBM Corp. 2019).

4. FINDINGS

To ensure a high level of reliability and minimum level of bias, the same teachers conducted and corrected the test and retest for each module. The descriptive statistics of the test for reliability of the four modules exams are displayed in Table (1). The maximum average grade was 17 for the test and 18 for the retest, respectively, while their minimum score was nil, with a standard deviation of 3.66 and 3.63 for the test and retest, respectively. In addition, their correlation coefficient was estimated to be $r^2 = 0.96$ and a Cronbach α of 0.98.

Table 1: Descriptive statistics of the test and retest scores on the four modules

	Min	Max	Mean	SD
Test	0.00	17.00	8.80	3.66
Retest	0.00	18.00	9.22	3.63

Furthermore, ANOVA analysis showed that the p-values for inter-students and test-retest were highly significant (p= 0.0001) (Table 2).

Table 2: Variance analysis of differences among students and between test and retest variables.

	D.F.	Sum of squares	Mean squares	Observed F	P value
Inter-students	1271	33162.621	26.092	49.676	< **0.001**
Error	1272	779.545	0.613		
Test-retest	1	111.964	111.964	213.166	< **0.001**
Error	1271	667.581	0.525		
Total	2543	33942.166	13.347		

Note: DF= Degree of freedom

These differences between the test and retest were also highly significantly affected (p value < 0.0001) by module, faculty, and their interaction (Table 3).

Table 3: ANOVA analysis of faculty, modules, and their interaction effect on the difference between test-retest scores over the four modules.

Source	Numerator degree of freedom	Denominator degree of freedom	Observed F	P value
Faculty	2	1260	38.980	**0.0001**
Module	3	1260	17.200	**0.0001**
Faculty * Module	6	1260	16.483	**0.0001**

5. DISCUSSION

Reliability of assessment is crucial in education. However, using the best method to measure reliability in practice has always been a difficult task for teachers. This study administered four tests to the students of the first semester during the academic year 2022/2023 on two occasions at three faculties of Letters and Humanities under the same conditions. The results showed a significant difference between the scores of students within the same module (Inter-students) and a significant difference in students' scores between test and retest for the four modules. To get more insights into the source of this variability, the interaction between modules and faculty is depicted in Figure 1.

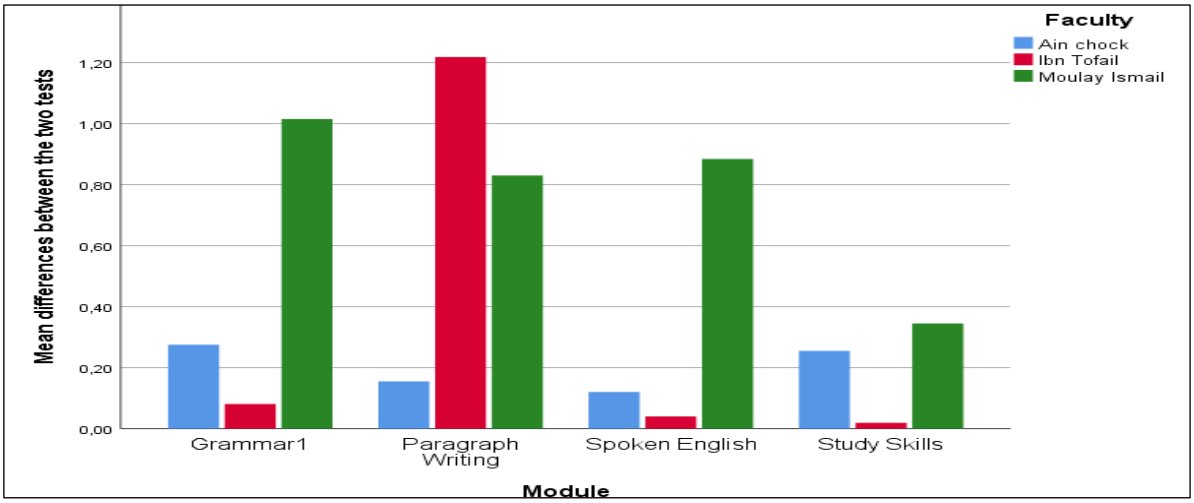

Figure 1: Test-retest mean differences (interaction between modules and faculties) at Ain Chock, Ibn Tofail, and Moulay Ismail Faculties of Letters and Humanities.

The highest mean differences between the two tests were at Moulay Ismail University. In contrast, less variability existed at Ain Chock and Ibn Tofail faculties. The description of the tests administered at Moulay Ismail and Ain Chock faculties revealed that these tests were similar and with almost the same degree of difficulty. Nevertheless, the variability of scores was much higher at Moulay Ismail Faculty because students scored more in the first test than in the second one (retest), for a variety of reasons. Students may not have expected to take the same test twice, so they did not search the topics in the tests before the second period. In addition, students might not have given much importance to the second test because they knew that it was administered for research purposes. In other contexts, test and retest scores may

slightly change because students learn from the test of the first period, or meanwhile, they improve their abilities (Alderson & Banerjee, 2002). Similarly, the test format may also help students become more familiar with the test and, therefore, score higher on the second test (Neukrug & Fawcett, 2015).

For paragraph writing scores in Figure 1, the mean difference of this test at Ibn Tofail University was the highest compared to the other paragraph writing and other tests in other faculties. It is worth mentioning that all the tests designed at Ibn Tofail Faculty contained multiple-choice items. These items provided several advantages when associated with this format. First, if written well, they are reliable because of only one possible answer. Second, they can be helpful at various educational levels (Coombe et al., 2007). Besides, scoring this item is quick, easy, and objective because different scorers will agree on the same scores for the same responses (McMillan, 2018). It also enables examiners to ask many questions covering many subject materials (Becher & Johnson, 1999) and helps students not to lose points for grammar accuracy, poor writing ability, or poor spelling (Zeidner, 1987). Besides, it is convenient for statistical analysis and can provide appropriate feedback for language teaching and learning (Brigui, 2017). Briefly, the literature on testing pedagogy confirms that students and faculty members prefer multiple-choice tests despite doing this for different reasons (Chan & Kennedy, 2002; Zeidner, 1987).

However, multiple-choice items have some drawbacks. First, this item is difficult to construct compared to constructed-response questions (Brown et al., 1997). In constructing multiple-choice tests, teachers may focus on removing ambiguous questions but find themselves designing examinations that are too easy and do not accurately indicate students' understanding (Simkin & Kuehler, 2005). Moreover, this type of test encourages guessing, which cannot be excluded in choosing the correct answer by students. The chance of guessing an answer to a question with four-answer options is 25%, and the chance will undoubtedly be higher if test takers can guess or eliminate one or two distractors (Brigui, 2017). Moreover, a multiple-choice test is an indirect test that cannot measure the natural language ability of the candidates. In this regard, the multiple-choice questions technique is invalid for two main reasons. First, people rarely use four selection ways to express understanding of a given topic. Second, they usually show their understanding of listening and reading by speaking and writing (Weir, 1990). Finally, multiple-choice tests favour rote learning and prevent students from organizing, synthesizing, and expressing knowledge in personal terms creatively (Brindgeman, 1992; Tuckman, 1993; Lukhele et al., 1994).

Analysis of the grammar test administered at Ibn Tofail faculty revealed that the questions were straightforward, and students were more likely to recall the test answers in the second period. The same was true for spoken English and study skills. The nature of the questions in the spoken English test designed by Ibn Tofail faculty made it easier for students to answer the test almost similarly in both periods because the items were short and well-constructed. Moreover, the suggested choices were not ambiguous, resulting in a high consistency in students' scores between test and retest.

The study skills test at Ibn Tofail faculty was the most reliable among all the tests administered among the three faculties. An insightful analysis of this test showed that its items were easy and not complicated. The nature of the questions made it easier for students to recall the

answers from the first test and led to the highest level of consistency. The quality of multiple-choice questions affects test reliability; good items increase reliability, while bad ones reduce it (Brigui, 2017). Consequently, the study skills test administered at Ibn Tofail faculty was the most reliable of all the tests at the three faculties studied.

The paragraph-writing test identified the highest variability between the test and retest administered to students at Ibn Tofail University (Figure 1). The test content and the objectives of teaching paragraph writing revealed that none of the items included in this test helped students write a paragraph. Unfortunately, that was the core objective of this module. To be valid, test designers should know that a writing test should only reveal writing ability and not test other skills (Richards, 2015). Unfortunately, this is not the case in the paragraph-writing tests designed for students at Ibn Tofail faculty. All the items and questions in this test assessed students' background knowledge but did not measure their ability to write a paragraph. These limitations in test design affected variability between test and retest. Students scored better in the first period than in the second one.

The irregularity between test and retest may also be due to many other factors. Test unreliability may emerge from test instructions, personal factors, and test scoring (Heaton, 1990). Similarly, measurement errors may come from three factors: examinee-specific, test-specific, and scoring factors (Brigui, 2017). In this regard, and since the paragraph-writing test administered at Ibn Tofail University adopted the multiple-choice format, this may be the source of unreliability. Multiple-choice tests should have several characteristics that guard against unreliability. Therefore, items must be evenly difficult, and test designers have to distribute evenly distractors to make the test reliable (Brown & Abeywickrama, 2010). In addition, test-specific factors may have caused the large variability between the first and the second paragraph writing test scores. The test was long and contained long tasks that required more concentration, which explains students' disinterest in the second-period tests. Moreover, the conditions in which the second test was administered might have been the source of variability in test scores. Students may not have had enough time to finish the test, or they may have felt tired.

Because of the factors that lead to test-retest unreliability, some scholars do not recommend using this method to determine the reliability of language tests. This method is often impractical and not frequently recommended (Heaton, 1990). However, the test-retest method is common in different fields to measure test reliability, especially in health and nursing. Moreover, the test-retest assessed the reliability of international tests such as TOEFL and Duolingo. Henning (1993) conducted test-retest analyses of the 'English Test as a Foreign Language'; test-length-adjusted reliability estimates were adequately high across reported components and total test scores, with raw test-retest coefficients ranging from 0.87 to 0.98. On the other hand, Settles (2016) reported several reliability measures for the first operational year of Duolingo. The results showed that the standard error of measurement was stable across the score range, the reliability and internal consistency coefficient were both 0.96 and the test-retest reliability coefficient was 0.84. Our coefficient was stronger and reached 0.96.

6. CONCLUSION

As one of the most influential methods of testing the reliability of language tests, the test-retest method needs to be investigated at a more comprehensive level. Most researchers and scholars have dealt with reliability at the theoretical level; however, a need for such studies is more than

welcome, especially under the context of these Moroccan faculties. Reliability is a significant criterion contributing to language tests' practicality and usefulness. Therefore, reliability testing may improve the quality of language tests. In this regard, the present study filled the gap mentioned in the introduction. Despite the difficulties and challenges encountered, we did measure the reliability of the first semester tests using the test-retest method. The results of this study revealed significant variability in scores not only between students within each module but also between the four modules administered at the three faculties. This has important implications for the curriculum designers, test designers, and educational officials in the Moroccan University Departments of English Studies. Curriculum designers should ensure that students in the English studies departments study the same content and work to achieve the same objectives. In addition, test designers should respect language assessment criteria when designing tests for common core students. Developing a test specification model is highly recommended in further studies to ensure that the tests designed for Moroccan common core students at tertiary education reflect language assessment principles.

REFERENCES

Abouabdelkader, S. (2018). *Moroccan EFL university students' composing skills in the balance: Assessment procedures and outcomes*. In Abdelhamid, A., & Abouabdelkader, H. (2018). *Assessing EFL writing in the 21st-century arab world: Revealing the unknown*. Gewerbestrasse: Springer.

Alderson, C., & Hamp-Lyons, L. (1996). *TOEFL preparation courses: A study of washback. language testing* 13(3), 280–297.

Alderson, C.J., & Banerjee, J. (2002). Language testing and assessment (Part 2). Language Teaching, 35, 79-113.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests. Oxford*: Oxford University Press.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. Assessment in Education: Principles, Policy & Practice, 18(3), 279-293.

Becker, W. E., & Johnston, C. (1999). *The relationship between multiple choice and essay response questions in assessing economics understanding.* Economic Record, 75(231), 348–357.

Bridgeman, B. (1992). *A comparison of quantitative questions in open-ended and multiple-choice formats*. Journal of Educational Measurement, 29, 253–271.

Brigui, H. (2017). *Investigating the ITU semi-computerized MCQ model: A systematic assessment of EFL students' attitudes and testing behaviors in the newly adopted e-testing form*. IOSR Journal of Research & Method in Education (IOSR-JRME). 7(3), 19-25.

Brown, D. B. (1996). *Testing in language programs*. New Jersey: Prentice Hall, Inc.

Brown, G., Bull, J., & Pendlebury, M. (1997). *Assessing student learning in higher education.* London: Routledge.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. New York: Pearson Longman.

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom Practices* (3rd ed.). London: Pearson.

Chan, N., & Kenedy, P. E. (2002). *Are multiple-choice exams easier for economics students? A comparison of multiple choice and equivalent constructed response exam questions*. Southern Economic Journal, 68(4), 957–971.

Chiedu, R. E., & Omenogor, H. D. (2014). *The Concept of Reliability in Language Testing: Issues and Solutions.* Journal of Resourcefulness and Distinction, 8(1).

Coombe, C., Folse, K., & Hubley, N. (2007). *A practical guide to assessing English language learners. Ann Arbor*: The University of Michigan Press.

Erford, T. B., & Bardhoshi, G. (2017). *Processes and procedures for estimating score reliability and precision, measurement and evaluation in counseling and development*, 50(4), 256-263.

Flutcher, G., & Davidson, F. (2007). *Language testing and assessment: An advance resource book.* London and New York: Routledge Taylor and Francis Group.

Ghaicha, A., (2016). *Theoretical framework for educational assessment: A synoptic review*. Journal of Education and Practice, 7(24).

Gipps, C. V. (1994). Beyond testing: *Towards a theory of educational assessment*. Washington, D.C.: The Flamer Press.

Gronlund, N. E. (1977). *Constructing achievement tests (2nd ed.)*. Englewood Cliffs: Prentice-Hall.

Gronlund, N. E. (1982). *Constructing achievement tests (3rd ed.)*. New Jersey: Prentice-Hall.

Harris, D. P. (1969). *Testing English as a second language*. New York: McGraw-Hill Book Company.

Heaton, J. B. (1990). *Writing English language tests*. New York: Longman Inc.

Henning, G. (1993). *Test-retest analyses of the test of English as a foreign language. TOEFL Research Reports (Report 45)*. Distributed by ERIC Clearinghouse.

Hughes, A. (1989). *Testing for language teachers (1st ed.)*. Cambridge: Cambridge University Press.

Hughes, A. (2003). *Testing for language teachers (2nd ed.)*. Cambridge: Cambridge University Press.

IBM Corp. Released 2019. IBM SPSS statistics for Windows, Version 26.0. Armonk, NY: IBM Corp.

Lukhele, R., Thissen, D., & Wainer, H. (1994). *On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests*. Journal of Educational Measurement, 31(3), 234–250.

McMillan, J. H. (2018). *Classroom assessment: Principles and practice that Enhance Student Learning and Motivation (7th ed.)*. New York: Pearson.

Maduekwe, A.N. (2007). *Principles and Practice of Teaching English as a Second Language.* Lagos: Vitaman Educational Books.

Neukrug, E. S., & Fawcett, R. C. (2015). *The essentials of testing and assessment: A* practical *guide to counselors, social workers, and psychologists (3rd ed.)*. Stamford: Cengage Learning.

Popham, J.W. (2007). Classroom Assessment: What teachers need to know. Boston: Pearson Education Ltd.

Richards, J. C. (2015). Key issues in language teaching. Cambridge: Cambridge University Press.

Sattler. J. M. (2001). *Assessment of children cognitive applications (4th ed.)*. USA: Publisher Inc.

Settles, B. (2016). *The reliability of Duolingo English test scores. Duolingo Research Report* DRR-16-02. englishtest.duolingo.com/resources.

Simkin, M.G., & Kuechler, W. L. (2005). *Multiple-choice tests and student understanding: What is the connection?* Decision Sciences Journal of Innovative Education, 5(1).

Tuckman, B. W. (1993). *The essay test: A look at the advantages and disadvantages.* NASSP-Bulletin, 77(555), 20–26.

Wells, C., & Wollack J. (2003). *An Instructor's guide to understanding test reliability, testing & evaluation Services*. Madison: University of Wisconsin.

Weir, C. J. (1990). *Communicative language testing*. New York: Prentice Hall.

Zeidner, M. (1987). *Essay versus multiple-choice type classroom exams: The student's perspective.* Journal of Educational Research, 80(6), 352–358.

About the Author

Mr. Anouar Mohamed El kasri ***is a high-school teacher of English and a doctoral candidate. He got his master's degree in 'Applied Language Studies' from Moulay Ismail University, Meknes in 2020 and enrolled in the applied Linguistics doctoral program at the same university in 2021. He is a member of the ELT Society (English Language Testing Society) and he has participated in many national and international conferences. He is interested in the fields of TEFL, Assessment design and testing, and language development.***

Dr. Mohammed Larouz ***is the Dean of the faculty of arts and humanities, at Moulay Ismail University, Meknes, Morocco, where he has worked since 2005. He also directs the PhD Program in Applied Linguistics and chairs the Research Group on Applied Linguistics & Language Development (ALLD) at the same faculty. He investigates applied linguistics questions in Morocco and has research interests in the fields of TEFL, sociolinguistics, research methodology, and communication.***

Dr. Moulay Sadik Maliki ***was born in Errachidia, Morocco. He has been a professor at the English Studies Department at the Faculty of Arts and Humanities, Hassan II University, Casablanca, since 1988. He has published many books and articles on intercultural communication, cultural and linguistic diversity in Morocco, culture and reading, inter alia. He has also participated in many national and international conferences. Dr. Maliki is interested in cultural issues in communication, applied linguistics, and translation.***

Dr. El Yousfi ***is a former Scientist at Plant Pathology Lab, National Institute of Agricultural Research (INRA), Morocco. He got his Engineer's degree in phytopathology from Minnesota University, USA in 1995 before he got his PhD in phytopathology from Hassan II University, Rabat, Morocco. He worked for INRA from 1985 to 2017 as an engineer. Dr. EL Yousfi published more than 30 peer-reviewed papers and served as an expert for several projects and organizations.***