International Journal of English Linguistics; Vol. 12, No. 3; 2022
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

Evaluating and Testing English Language Skills: Benchmarking the TOEFL and IELTS Tests

Haytham Bakri[1]

[1] College of Languages and Translation, King Saud University, Saudi Arabia

Correspondence: Haytham Bakri, College of Languages and Translation, King Saud University, P.O. Box 87907 Riyadh 11652, Saudi Arabia.

Received: March 3, 2022 Accepted: April 10, 2022 Online Published: April 19, 2022

doi:10.5539/ijel.v12n3p99 URL: https://doi.org/10.5539/ijel.v12n3p99

Abstract

Testing English Language skills cannot be ignored in English Language classrooms all over the world. Most importantly, it is pertinent to describe how students view their own achievements. Reports have repeatedly shown that students' grades often differ from their expectations. Standardized English tests are an important requirement for international students. TOEFL and IELTS are two set of tests that are widely used worldwide. Hence, this study aimed to test the validity of placement tests (TOEFL and IELTS). To achieve the objective of the study, data was gathered on the face validity and construct validity of TOEFL and IELTS exams from respondents who were taking the exams in Riyadh area of Saudi Arabia. A total of 60 students participated in the study by filling the questionnaire. Data gathered was analyzed using SPSS. The results of the study were presented in tables and figures. The tests' reliability was determined using the Rasch model. The analysis showed that both tests were valid at r-score = (.477; .288; .183; .012) for reading, listening, speaking, and writing skills, respectively. The data analysis revealed that the placement tests chosen by students at the center (TOEFL and IELTS) were valid and reliable. The analyses conducted showed that Reading (0.291266), Speaking (0.343007), Listening (0.567623) and Writing (0.35101) skills constructed against face validity were valid, (between -1.0 to 1.0). This was proven by the Pearson Product Moment Correlation. The author concluded that the assessment of the tests' validity and reliability showed that the placement test instruments were dependable as well as valid, and the test takers face validity assessment provided evidence of the tests' effectiveness.

Keywords: IELTS, TOEFL, language tests, validity, reliability

1. Introduction

Testing English Language skills cannot be ignored in English Language classrooms all over the world. Most importantly, it is pertinent to describe how students view their own achievements. Reports have repeatedly shown that students' grades often differ from their expectation (Bacha, 2002). Hence, English Language skills testing is not rigid but flexible. In testing English Language skills tests, the keywords are **validity** and **reliability**. Particularly, this body of research has generated such pertinent questions as: how is the process of English Language testing evaluated? What are the acceptable standards by institutions? What methods have researchers found acceptable? What methods have ELT teachers found to be reliable for testing English Language skills?

It is, therefore, pertinent for academic institutions to make a thorough meticulously evaluation of a particular language test before giving to the test takers. The researcher has discussed in this paper that there are various ways that can be undertaken to improve language tests, and apart from its few flaws IELTS and TOEFL are still two of the most authentic, effective and widely accepted English language proficiency tests (Peltekov, 2021). These tests are obviously significant language proficiency testing tools and a kickstart for the improvement and measurement of English language skills. The present paper is a veritable attempt to show the validity of those tests. This paper, to say the least, is expected to help language teachers and test takers "to better understand the structure of the test and to reflect on its usefulness in a more informed and objective way" (Peltekov, 2021, p. 395). There are many factors to consider when assessing and evaluating a testing method in English Language. They are designed to meet the diagnostic and proficiency needs of learners. Most of the processes need however, careful evaluation. That is, they are based on score grading. Hawkey (2004) described the following methods to teach children how organize historical information:

1) The pre-scientific method of testing (grammar-translation method)

2) The psychometric structuralist testing method (direct, reading, and audio-lingual)

The pre-scientific (grammar-translation method) method of language testing uses the following approaches:

a. translating passages from L1 (first language) to TL (target language); and TL to L1.

b. assigning comprehension passages with L1 questions.

c. requesting essay writing in TL

d. asking questions in L1 about books that have been studied

The psychometric structuralist method of language testing uses the following approaches:

a. test of grammar using the multiple-choice questions, one-word completions, one-word transformation.

b. test of lexis.

c. test of comprehension (using multiple-choice, true/false, and sentence completion);

d. reading aloud; and

e. writing tests at sentence level (completions, conversion, and sequencing)

In his Ph.D. thesis on relationship between TOEFL and IELTS, Arcuino (2013) laid the claim that there was no significant difference between placement tests (TOEFL and IELTS) in association with academic success. Standardized English tests are an important requirement for international students. TOEFL and IELTS are two tests that are widely used worldwide (Arcuino, 2013) to evaluate the levels of language skills. The English Language department plays the role of support for acquisition of English Language skills as well as content study in universities. Before teachers are granted the role of English for Academic Purposes (EAP) instructors, they are expected to have attained an MA in Linguistics and written the TESOL exam. The purpose of this is to ensure that they can teach and test students accordingly. In testing, it is expected that questions are set to meet the essential criteria. These include:

1) The questions are kept to a minimum in length and content.

2) They should, test more than one-construct—testing only one construct would be considered invalid on prima-facie grounds.

Since TOEFL and IELTS are important in determining international students' admission application, it becomes pertinent that validity tests be conducted on these tests. This study joins a body of research that evaluates the validity of placement tests. Many other researchers have assessed the validity as well as the reliability of TOEFL and/or IELTS. Constant analysis of placement test is important to ensure consistency of quality. Consequently, this study gathered data on the ***face validity*** of TOEFL and IELTS exams based on the responses of participants taking the tests in Riyadh area of Saudi Arabia.

2. Literature Review

2.1 Reliability of English Language Skills Testing

Fulcher (1997) authored a paper titled, "An English Language Placement Test: Reliability and Validity", in it, he opined that, "Fair and accurate assessment of student abilities, and referring individuals to appropriate language support courses (the in-sessional program), is an essential support service to all academic departments" (p. 113). This study was associated with the University of Surrey. The skill testing taken by all students was a means of picking out students who required English Language Writing skill support to succeed. Incoming students were evaluated irrespective of their native language. They were also not discriminated based on their subject of specialization. By doing this, the university was able to ensure that the number of students who would not fit into learning in the university system would be minimal. This assessment is important in realizing this goal of the university. Hence, it became important that assessment process be evaluated for reliability and validity. ***Face validity*** refers to the face value of the test, it addresses such questions as how the test appears to students and non-stakeholders. This facet of the test evaluation dictates that tests be connected to students' actual writing needs. ***Content validity*** deals with the content that students are supposed to learn. Effective writing tests are deeply based on methods, and they should measure the content (Raharjo, 2020).

Earlier researchers have also pointed out that TOEFL and IELTS are important and useful for nonnative speakers of English who either wish to study abroad or appear for an interview to join a job. They help test takers to know their English language proficiency level in a proper, practical, and authentic way, and encourage them to make their personal efforts to improve themselves professionally (Renandya et al., 2018; Richards, 2017). Wulyani et

al. (2019) have observed that teachers and language planners must prepare a suitable language teaching syllabus so that language learners are motivated to continue their learning and take the TOEFL and IELTS tests to improve and verify their language proficiency level for their professional development. They also suggest programs for teachers to prepare students to take those tests. These studies have concentrated on changes to be undertaken by institutions as regards curriculum development and teaching methodology. Keeping in tune with those suggestions, the present researcher has attempted to explore and analyze the efficacy of the piloted language instructions and the opinions of test takers themselves with regard to the efficacy of the two tests.

2.2 Placement Tests

Placement tests are widely used and accepted for testing the levels of language skills. Usually, in assessing English Language, placement tests are used with language-minority students, students with less-privileged language opportunities. In TOEFL iBT, the writing task is based on reading, listening and writing. Language tests should reflect how a language is used. It reflects real-life authenticity. Construct demonstrates the difference between what is being evaluated and what the test hopes to assess. For example, writing constructs are specific to the demand of pronunciation and writing styles. In class, teachers can measure such constructs as writing ability and progress. Ideally, writing tests addresses multiple facets of English Language. This was the goal of the TOEFL iBT implementation. The ideal writing theory reveals students' performance on spelling, punctuation, and grammar. Another aspect of testing is the criterion-related testing which refers to how test measures compare on an external level. "The first aspect of criterion-related validity is concurrent validity, which is the extent to which the results of the question test agree with another independent, highly dependable second assessment method" (Hughes, 1989, p. 23). It relates to whether such tests are compatible with or effective among other language writing tests. The interactivity level in a test requires test takers to demonstrate linguistic knowledge and strategic competence (Luo, 2015). Language Tests are always envisioned to attend to the requirements of an educational system or a society at large (Bachman, 2000). As such, tests such as TOEFL and IELTS are referred to as placement tests rather than competency tests. Such factors considered in assessment tests include ***face validity*** (that is, students' assessment/perception) of the test; and ***content validity*** (gaining knowledge on whether placement tests evaluate course content). The method used by Fulcher (1997) in the Surrey University reliability and validity test expanded on the method described by Alderson (1991). Both methods concentrated on four areas:

1) Using pool judgement to determine cut scores for placement

2) Using statistic means to analyze data for the test

3) Developing parallel forms; and

4) Investigating face validity with questionnaire

3. Research Design

The study sought the participation of English Learners who were about to take the TOEFL and IELTS exams. This study gathered data on the ***face validity*** and construct validity of TOEFL and IELTS exams according to respondents who were going to take these tests in the Riyadh area of Saudi Arabia. Hence, the study compiled the responses of respondents as regards their perspective on the tests they were about to take. Some students were, however, going to take the test for a second time, that is, they were repeaters, because they had failed previous tests or had their results expired. Hence, the questionnaire contained information on the perspective of respondents on the tests they have taken or were about to take. The ***construct validity*** aspect of the questionnaire assessed important themes of the test by measuring the likelihood of the sentences below on a linear scale (Biber & Gray, 2013; Chen & Sheehan, 2015):

1) The content of the test is relevant to and representative of the kinds of tasks and written and oral texts that students encounter in college and university settings.

2) Tasks and scoring criteria are appropriate for obtaining evidence of test takers' academic language abilities.

3) Academic language proficiency is revealed by the linguistic knowledge, processes, and strategies test takers use to respond to test tasks.

4) The structure of the test is consistent with theoretical views of the relationships among English language skills.

5) Performance on the test is related to other indicators or criteria of academic language proficiency.

6) The test results are used appropriately and have positive consequences.

A total of sixty students participated in the study by filling the questionnaire. Data gathered was analyzed using SPSS. The results of the study were presented in tables and figures. Data was also analyzed using inferential statistics. Relative inferences were made, and the results were explained in the Discussion of the study. The concluding part of the research paper contains information on the recommendations for future research.

4. Results and Discussion

The following Table 1 shows that students were presented with a questionnaire to assess their English Language proficiency levels, to determine how much of the skills they had learnt, and their language weaknesses and strength. These questionnaires also sought to help language teachers to work on other test samples and questionnaires to analyze further the efficacy of the tests and the measure to be undertaken to qualify them.

Table 1. Students' success in the exams

	Yes	No	Total
Success in the exams	48 (80%)	12 (12%)	**60 (100%)**
Requirements met	36(60%)	24(40%)	**60 (100%)**

Responding to questions on whether they passed previous placement tests (IELTS and TOEFL), 20% of the respondents mentioned that they have not been successful at their former attempts, majority (80%) recorded that they were successful in their attempts. On whether their score tests meet the requirements for what they needed it for, 60% mentioned that it fulfilled their requirements, the remaining 40% mentioned that they will be taking the tests in the future to meet their requirements (Table 1).

This paper evaluates the general construct validity of such tests based on the responses and reactions gathered by sixty students from the Riyadh region of Saudi Arabia. It is pertinent to note here that in their study on a similar kind of test evaluation regarding the German language, Eckes and Grotjahn (2006) measured the construct validity of C-tests, which is a language proficiency test. Although testing a different language (the German language), the C-test is supposedly designed to test proficiency. In the study undertaken by Eckes and Grotjahn (2006), a total of 843 participants from 4 study centers took the tests. To measure the construct validity of the test, the Rasch measurement modelling was used. It provided unambiguous evidence as to whether the test was highly dependable (reliability). Much like English language, it evaluated the reading, listening, speaking and writing skills of individuals. Conclusively, a closer look at the Rasch measurement modelling technique revealed the trustworthiness of test score meaning and its interpretation. As regards validity, any modifications have occurred in the definition of validity. Validity is the degree to which a test measures what it claims to measure. Looking at the efficacy, accuracy, and reliability of the Rasch measurement model as used by the aforementioned German language researchers, this measurement model was used in the present study to establish the instrument's reliability, as it provides a clear indicator of each item's reliability.

The following were the average scores in Reading, Listening, Speaking and Writing skills for test takers who had formerly partaken in placement tests such as IELTS and TOEFL. These test scores are based on a total of 100.

Table 2. Test scores of respondents

	Average	Total Possible	Minimum Score	Maximum Score
Reading	75.83	100	37	71
Listening	65.97	100	41	73
Speaking	60.37	100	37	81
Writing	49.3	100	35	63

In evaluating the reliability and validity of placement tests, Fulcher (1997) revealed the essence of testing placement tests. Since these tests are important in determining international students' admission application, it becomes pertinent that validity and reliability test be carried out on them. Although many other researchers have over time, conducted tested the validity and reliability of TOEFL and IELTS, there is still need for constant analysis to ensure consistency of quality. Hence this study aimed to determine the validity and reliability of placement tests (TOEFL and IELTS).

Findings from this research revealed that the placement tests were valid as well as dependable, majority of respondents also revealed that these tests fulfilled their requirements to study abroad. Eckes and Grotjahn (2006)

measured the construct validity of c- tests (language proficiency tests) to assess the writing, speaking, listening, and reading skills using the Rasch measurement, similarly in TOEFL iBT, the writing task is based on reading, listening, and writing because language tests should reflect how a language is used. Interpreting the Rasch reliability value, the Rasch scores follow the expected response pattern (-2.0 to 2.0).

Although Arcuino (2013) in his thesis claimed that there was no significant difference between placement tests TOEFL and IELTS in association with academic success, findings from this study have revealed that the validity and reliability of these placement tests and how it goes a long way in determining the academic capacity of foreign students.

4.1 Reliability Value (r-value) of English Skills Test Scores

Subjects have a higher probability of answering easier items correctly and a lower probability of answering difficult items correctly. This assumption is based on prior knowledge of existing literature. According to the Rasch Model, the log of the odds of success was taken to calculate the probability of students' ability. This is the phenomenon of the Item Response Theory. It is based on the assumption that latent traits are quantified based on fundamental assumptions that a subject's response to an item is a function of the differences between his abilities and the characteristics of the item.

Rasch model = log(p/1-p)

Where p is the percent score.

The r-values for these test scores are presented in the next table.

Table 3. Reliability: R-values of tests scores using the Rasch Model

	Average	Minimum Score	Maximum Score	r-Score
Reading	75.83	37	71	.477
Listening	65.97	41	73	.288
Speaking	60.37	37	81	.183
Writing	49.30	35	63	.012

All of these values show that the test of the individual for reading, listening, speaking and writing skills are reliable. Interpreting the Rasch reliability value, the Rasch scores follow the expected response pattern (-2.0 to 2.0). Test scores that fall outside of these white paths are not considered reliable. All of the test scores above **(.477; .288; .183; .012)** are reliable. It thus shows that the placement tests chosen by students in these study centers (IELTS and TOEFL) are reliable.

4.2 Construct Validity of Placement Tests

Correlation coefficients are used to determine the validity of test constructs. For this study sample, each of the English Language skills tests scores (averages) – Reading, Listening, Speaking and Writing was compared against the items in the construct validity section of the questionnaire using the Pearson Product Moment Correlation (PPMC). These items were recorded as face validity as they recorded the opinions of individual test-takers.

Table 4. Average scores of test takers' skills

	Average Score
Reading	75.83
Listening	65.97
Speaking	60.37
Writing	49.3

The figure below represents the face validity scores. The face validity score was measured on a five-point Likert scale. There were 6 items in total. The table below represents that averaged face validity scores as recorded by respondents.

Table 5. Averaged face validity scores of respondents on placement exams (IELTS, TOEFL)

Face Validity Construct	Score
The content of the test is relevant to and representative of the kinds of tasks and written and oral texts that students encounter in college and university settings.	**3.50**
Tasks and scoring criteria are appropriate for obtaining evidence of test takers' academic language abilities.	**4.00**
Academic language proficiency is revealed by the linguistic knowledge, processes, and strategies test takers use to respond to test tasks.	**4.16**
The structure of the test is consistent with theoretical views of the relationships among English language skills.	**4.00**
Performance on the test is related to other indicators or criteria of academic language proficiency.	**2.50**
The test results are used appropriately and have positive consequences.	**3.50**

4.3 The Pearson Product-Moment Correlation of Reading, Listening, Speaking and Writing Skills Score Set against Construct Validity items

The correlation tests run for English Language Speaking test score values against the face construct validity revealed the following correlation scores: Pearson correlation for Reading skill (Learners' ability) against face validity to determine construct validity – general validity of the test = *0.291266*. Pearson correlation for Listening skill (Learners' ability) against face validity to determine construct validity – general validity of the test = *0.567623*. Pearson correlation for Speaking skill (Learners' ability) against face validity to determine construct validity – general validity of the test = *0.343007*. Pearson correlation for Writing skill (Learners' ability) against face validity to determine construct validity – general validity of the test = *0.35101*.

All the above correlation scores are significant. Correlation values are significant at -1.0 to 1.0. All of the above values for English Language skills against the face validity construct fall within these range of values (Reading: 0.291266; Listening: 0.567623; Speaking: 0.343007; Writing: 0.35101). These correlation result shows a positive value between English Language skills and the face validity scores. Expansively, it proves the validity of the placement tests (IELTS and TOEFL) taken by students at the English learning center.

Several researchers have repeatedly used the Rasch model score to test for the reliability of language placement scores. While numerous studies have corroborated these research findings (Mokshein, Ishak, & Ahmad, 2019) which proved the validity of placement tests such as TOEFL and IELTS, a few researchers have noted that this validity and reliability does not mean that test takers are not distracted in certain ways. As such many researchers have concluded that overgeneralization should be checked to ensure consistent reliability and validity of these tests (Runnels & Bunkyo, 2012).

5. Conclusion

Many studies have investigated placement tests (IELTS, TOEFL validity and reliability) as efficient decision-making tools in language proficiency for students' placement or career roles (Karjo & Ronaldo, 2019). Such studies evaluate the content, construct, and predictive validity. In this study, the author assessed the test takers ability as well as the face validity of the tests. Assessment of the tests' validity and reliability showed that the placement test instruments were dependable as well as valid. The test takers face validity assessment provided evidence of the tests' effectiveness. Furthermore, Shohamy (2013) also argued that the investigation of the tests based on the test takers' experience is significant because it provides stakeholders, especially testers, with new understanding about tests and their meaning.

5.1 Limitation

This test viewed the placements tests assessed in this study as a group. Further studies could provide an insight into the individual test's (IELTS and TOEFL) validity and reliability.

References

Alderson, J. C. (1991). Language Testing in the 1990s: How Far Have We Come? How Much Further Have We to Go? In Current Developments in Language Testing. *Anthology Series*, *25*, 1–27. Washington, D.C.

Arcuino, C. L. T. (2013). *The Relationship between the Test of English as a Foreign Language (TOEFL), The International English Language Testing System (IELTS) Scores and Academic Success of International Master's Students*. Ph. D. thesis submitted at Colorado State University, Fort Collins, Colorado. http://hdl.handle.net/10217/78818.

Bacha, N. N. (2002). Testing Writing in the EFL Classroom: Student Expectations. *English Teaching Forum Journal*, *40*(2), 14–19.

Bachman, L. F. (2000). Modern Language Testing at the turn of the century: Assuring that what we counts. *Language Testing*, *17*(1), 1–42. https://doi.org/10.1191/026553200675041464

Biber, D., & Gray, B. (2013). Discourse Characteristics of Writing and Speaking Task Types on the TOEFL IBT Test: OF WRITING AND SPEAKING TASK TYPES ON THETOEFL iBT®TEST: A Lexico-Grammatical Analysis. *ETS Research Report Series*, *1*, 128. https://doi.org/10.1002/j.2333-8504.2013.tb02311.x

Chen, J., & Sheehan, K. M. (2015). Analyzing and Comparing Reading Stimulus Materials Across theTOEFL®Family of Assessments. *ETS Research Report Series*, *1*, 1–12. https://doi.org/10.1002/ets2.12055

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, *23*(3), 290–325. https://doi.org/10.1191/0265532206lt330oa

Fulcher, G. (1997). An English language placement test: Issues in reliability and validity. *Language Testing*, *14*(2), 113–139. https://doi.org/10.1177/026553229701400201.

Hawkey, R. (2004). *A Modular Approach to Testing English Language Skills: The Development of the Certificates in English Language Skills (CELS) examinations*. Cambridge: Cambridge University Press.

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

Karjo, C. H., & Ronaldo, D. (2019). *The Validity of TOEFL as entry and exit college requirements: Students' perception* (pp. 326–330). Conference: Proceedings of the Eleventh Conference on Applied Linguistics (CONAPLIN 2018). https://doi.org/10.2991/conaplin-18.2019.277

Luo, K. K. (2015). Validity Considerations in Designing a Writing Test. *Studies in Literature and Language*, *10*(5), 19–21.

Mokshein, S. E., Ishak, H., & Ahmad, H. (2019). The Use of RASCH Measurement Model in English Testing. *Jurnal Cakrawala Pendidikan*, *38*(1), 16–32. https://doi.org/10.21831/cp.v38i1.22750

Peltekov, P. (2021). The International English Language Testing System (IELTS): A Critical Review. *Journal of English Language Teaching and Linguistics* (JELTL), *6*(2), 395–406.

Raharjo, S. D. (2020). Students' Perception: Assessing English Competence in TOEFL As a Standardized English Language Proficiency Test in Indonesian's Higher Education. *Intensive Journal*, *3*(2), 40–48.

Renandya, W. A., Hamied, F. A., & Nurkamto, J. (2018). English language proficiency in Indonesia: Issues and prospects. *The Journal of Asia TEFL*, *15*(3), 618–629. https://doi.org/10.18823/asiatefl.2018.15.3.4.618

Richards, J. C. (2017). Teaching English through English: Proficiency, Pedagogy, and Performance. *RELC Journal*, *48*(1), 7–30. https://doi.org/10.1177/0033688217690059.

Runnels, J., & Bunkyo, H. (2012). Using the Rasch model to validate a multiple choice English achievement test. *International Journal of Language Studies*, *6*(4), 141–153.

Shohamy, E. (2013). The discourse of language testing as a tool for shaping national, global, and transnational identities. *Language and Intercultural Communication*, *13*(2), 225–236. https://doi.org/10.1080/14708477.2013.770868

Wulyani, A. N., Elgort, I., & Coxhead, A. (2019). Exploring EFL teachers' English language proficiency: Lessons from Indonesia. *Indonesian Journal of Applied Linguistics*, *9*(2), 263–274. https://doi.org/10.17509/ijal.v9i2.20217