languages

Article

Establishing Reliability and Validity of an Online Placement Test in an Omani Higher Education Institution

Samia Naqvi [1,*], Reema Srivastava [1], Tareq Al Damen [2], Asma Al Aufi [1], Amal Al Amri [1] and Suleiman Al Adawi [1]

[1] Centre for Foundation Studies, Middle East College, Muscat 124, Oman
[2] Centre for Preparatory Studies, Sultan Qaboos University, Muscat 123, Oman
* Correspondence: snaqvi@mec.edu.om

Abstract: Although placing students in the appropriate proficiency levels of post-secondary English programs is crucial for optimal learning, the evaluation of placement tests (PTs) in terms of establishing their reliability and validity is relatively under-researched. This study assesses the validity, reliability, and effectiveness of an in-house online PT. The content validity was established through the internal and external moderation of the question papers and answer keys while criterion-related concurrent validity was established via IELTS benchmarking. New Student Survey was used to investigate the face validity. The internal consistency and reliability of the reading test items were measured using Cronbach's alpha while descriptive statistics were calculated for the listening test. Paired sample *t*-test (dependent *t*-test) was used to assess the inter-rater reliability of the speaking and writing tests which were double-marked. The data analysis revealed that the PT was effective in placing students at different levels of the foundation program (FP) and the statistical analyses conducted to test the reliability and validity showed positive results for most of the test versions. The study offers useful insights to test developers and policymakers regarding the authentication of in-house tests and the creation of guidelines for PT design and evaluation.

Keywords: foundation program; higher education institution; Omani HEI; placement test; reliability; validity

Citation: Naqvi, Samia, Reema Srivastava, Tareq Al Damen, Asma Al Aufi, Amal Al Amri, and Suleiman Al Adawi. 2023. Establishing Reliability and Validity of an Online Placement Test in an Omani Higher Education Institution. *Languages* 8: 61. https://doi.org/10.3390/languages8010061

Academic Editors: Dina Tsagari and Henrik Bøhn

Received: 8 December 2022
Revised: 23 January 2023
Accepted: 26 January 2023
Published: 22 February 2023

1. Introduction

Due to the use of English as the medium of instruction in many universities around the world, including the Middle East, the standardization of in-house locally developed English placement tests (PTs) has gained substantial importance. Most of the students joining such universities are non-native speakers (NNS) who need to undertake a foundation program (FP) to develop their English language proficiency. Fair and accurate assessment of students' abilities and their placement into appropriate language courses in the FP, based on their language proficiency, is crucial for homogenous grouping and optimum teaching and learning (Fan and Jin 2020; Fulcher 1997; Fulcher et al. 2022; Hille and Cho 2020; Liao 2022; Shin and Lidster 2017).

Based on the specific requirements and other academic considerations, higher Education institutions (HEIs) either use commercially available tests or develop in-house tests to place students into different levels of the FP. It is believed that in-house tests ensure a range of benefits as they are customized to the specific curricular goals of the academic programs offered by the institutions (Chung et al. 2015) and are cost-effective (Jamieson et al. 2013). However, the effectiveness of such tests in placing students into appropriate levels is often questioned as they might suffer from validity and reliability issues (Fan and Jin 2020). An invalid and unreliable test tends to place students at the wrong levels, which may have an adverse impact on the student's proficiency and develop negative attitudes towards the university among students (Al-Adawi and Al-Balushi 2016). In addition, the teaching and learning process can be a struggle for both teachers and students when students are misplaced (Johnson and Riazi 2017). Inaccurate placement may have financial implications, impact students' degree plans, and lead to an adverse impact on their motivation levels (Hille and Cho 2020).

Due to the implications of PT results on score users, it is important to ensure that the test scores are accurate in informing placement decisions. By the same token, it is essential to establish the validity and reliability of the in-house developed PTs. However, there is surprisingly little research on the design, reliability, and validity of PTs, although they are perhaps among the most used measures within institutions (Wall et al. 1994).

PTs, in general, follow several methods to place students at different levels of English language programs and may include interviewing, essay writing, multiple-choice tests, or a combination of different methods. Therefore, the evaluation of their reliability and validity depends, to a large extent, on their specific characteristics (Shin and Lidster 2017).

This study assesses the validity and reliability of an in-house online PT designed by the Centre for Foundation Studies (CFS) in a private HEI in Oman. Broadly, the methodology used to ascertain the reliability and validity of the test in this study is similar to that of Fulcher (1997) and Wall et al. (1994), which is detailed in the literature review section. However, this study attempts to expand on their methodology for the evaluation of the four language skills assessed as part of the English PT designed by the CFS at the research site. The study contributes to the existing body of language assessment research concerning the validity and reliability of PTs in placing non-native English speakers into a language program. The methods used in the study can be used to ascertain the strength of assessments used by other institutions at both local and global levels.

2. Literature Review

2.1. In-House (Local) Versus Commercially Produced Large-Scale PTs

Several HEIs use commercial or standardized PTs for placing students in undergraduate programs, while many others design their own tests. Standardized PTs can be appealing for many reasons. First, they relieve universities from the stress of time constraints during the development and scoring of tests, especially when online tests can be taken at multiple locations by many candidates (Jamieson et al. 2013). Moreover, language programs also trust commercial/standardized PTs because of reliability issues with local PTs (Hilgers 2019). Despite these advantages, commercialized PTs cannot discriminate among students of varying proficiency levels (Westrick 2005). In-house PTs offer a range of advantages over commercial tests since they measure students' abilities within a specific institutional context (Westrick 2005) and can be customized according to specific curricular goals (Chung et al. 2015), whereas commercial PTs cannot be linked closely to any specific institution. According to Dimova et al. (2022), "While large-scale tests have a wide-reaching and often overwhelming impact, within generalized contexts, local language tests address specific needs and have a deeper influence on day-to-day language assessment practice and research" (p. 243). Thus, the development of customized PTs and their widespread use is a result of the practical need to assess English language learners' abilities locally (Fox 2009) which can be made possible via an in-house test.

2.2. Validity and Reliability Studies of Placement Tests

Interest in language testing-related issues has increased over time; however, "... validity/validation received the highest interest across periods" (Dong et al. 2022, p. 1). Moreover, the validity of a PT is critical for allowing a better understanding of the test scores and the consequences of placement decisions based on these scores (Chun 2011; Li 2015). Wall et al.'s (1994) study conducted at the University of Lancaster is the first one in the field of language testing that addressed the evaluation of placement instruments in depth. They investigated face validity (through a student survey), content validity (using teacher interview), construct validity (by measuring Pearson product–moment correlation coefficients), concurrent validity (with student self-assessments and subject and language tutors' assessments), and reliability by calculating mean and standard deviation (SD) from students' scores. They concluded that, overall, the PT content was satisfactory, the test balance was appropriate, and no students were reported to be wrongly placed in their classes. The limitation of their study was not finding external criteria to measure concurrent validity. Building on Wall et al.'s pioneering work, Fulcher (1997) conducted a reliability and validity study of the PT used at the University of Surrey. For the investigation of reliability, correlation coefficients, means, and SDs (inter- and intra-rater reliability) were established for rating patterns in the writing task. For structure and reading comprehension, a logistic model was used and Rasch analysis was performed. Both Wall et al. (1994) and Fulcher's (1997) studies used Pearson product–moment correlation for construct validity; however, Fulcher also used inter-rater reliability for writing assessment. His findings were similar to Wall et al.'s (1994) findings, where most of the students considered the test fair with a few of them voicing their concern regarding the ambiguity of some test items. Fulcher's addition to Wall's was the use of concurrent validation using TOEFL. In a subsequent study, Fulcher (1999) focused on the computerization of a PT and assessed the usefulness of the computer-based test (CBT) as a placement instrument by comparing it with the pencil-and-paper form of the test. This is a seminal study since this was the very first one conducted on computerizing PTs.

Similar to Fulcher (1997), Nakamura (2007) also performed a Rasch analysis to validate the in-house Reading PT used at the Faculty of Letters at Keio University. He used the item characteristic curve (ICC) for item analysis to establish construct validity and concluded that 94% of the test items fitted the model. Face validity was investigated using student questionnaires in both studies. Nakamura (2007) used the person separation index to investigate reliability, which is similar to the Cronbach alpha. The reliability of the test had a score of 0.78 which established that the items in this test were internally consistent. Kim and Shin (2006) also assessed the construct validity of the multiple-choice test using the Pearson product–moment procedure to determine the correlation between the different domains of the reading (gist, vocabulary, inference, and detail) and writing (content, organization, and form) tasks. To estimate the internal consistency reliability of the multiple choice items of the reading test, Cronbach's alpha was calculated. Even though their study details the process of PT design, evaluation, and analysis, the limited number of items and sample size affected the reliability estimate. Kim and Kim's (2017) approach to validation of the English PT used at Kyun Hee University can also be considered similar to the studies mentioned above. The internal consistency and reliability of the test items measured using Cronbach's alpha were 0.89, indicating the high reliability of the test items. The outcome of the classical test theory method showed the item difficulty of 0.48 and item discrimination of 0.448. However, their PT only considered the receptive skills of reading and listening for placing students.

Messick's (1996) unified theory of test validity and Kane's (2013) argument-based approach have also been used for the validation of PTs. Li (2015) used the self-assessment tool within an argument-based validity framework (Kane 2013) to validate the PT used at a Midwestern university. He also employed the Rasch-based item analysis (Fulcher 1999; Nakamura 2007). The results revealed that the self-assessment items had acceptable reliabilities and item discrimination; however, the multivariate–multimethod analysis revealed weak to moderate correlation coefficients between the candidates' self-assessments and their performances on the PT and TOEFL IBT. Huang et al. (2020) combined Messick's and Kane's approaches to validate the speaking test used in their institution. Significant relationships between speaking test scores, self-ratings of speaking skills, and instructors' end-semester exam ratings were observed. Yet, there were some issues with rubric design and limited training in terms of test administration and scoring. Limited assessment literacy is a concern raised by other researchers also in the field of language testing (for example, Ashraf and Zolfaghari 2018; Coombe et al. 2020; Genç et al. 2020). It is important to note that Huang et al.'s (2020) study considered only speaking scores in placing students which is not sufficient for appropriate placement in ESL programs. A more structured approach for speaking assessment based on the Common European Framework of Reference (CEFR), especially in the case of large groups, is suggested by Emperador-Garnace (2021).

The use of standardized tests n placing students is acceptable (Jamieson et al. 2013; Hilgers 2019) yet a debatable practice in terms of placing students into exact levels of ESL programs. Liskinasih and Lutviana (2016) compared students' TOEFL scores with final test scores using Pearson product–moment correlation and found a moderate positive correlation level (0.41). The bivariate correlational analysis revealed a positive correlation (r = 0.643) between scores of the listening component of the TOEIC and the sentence repetition placement test in a study conducted by Topor (2014) on Japanese learners. Liao (2022) investigated the accuracy and validity of placement decisions based on the English GSAT scores of Taiwanese university students, with a focus on its associations with the General English Proficiency Test (GEPT) and students' performance in the course. The GSAT was reported to have appropriately placed lower or higher-level students in EFL classes but did not distinguish well for the borderline cohort.

As opposed to other researchers (Jamieson et al. 2013; Hilgers 2019; Liao 2022; Topor 2014), Kokhan (2013) is against the idea of placing students in ESL programs based on standardized test scores. He examined the validity of SAT, ACT, and TOEFL iBT scores as a substitute for the English PT and concluded that there is a 40% probability that most prospective students might be placed at the wrong level. This argument adds value to the importance of an in-house test that is aligned with the ESL curriculum. Nakamura (2007) also asserts that the content, level, and purpose of standardized tests are not suitable for placing students.

In the Middle Eastern context, the research evidence on the validation of in-house tests is very limited (Al-Adawi and Al-Balushi 2016; Mahfoud 2021; Rahal and Dimashkie 2020; Rouhani 2008). Rahal and Dimashkie (2020) updated a customized English PT used at an American university in the Middle East to improve its security, reliability, and validity. They created a new test bank, revised the grading rubric, and then created a test specifications document. They call the process Creational Reverse Engineering. Rouhani (2008) administered the Michigan Test of English Language Proficiency (MTELP) and an in-house C-Test to 144 Iranian university-level students. The results revealed fairly high criterion-related validity, high reliability, and acceptable content relevance of the C- test. The extracts used in the C-Test turned out to measure similar attributes as the MTELP, showing significant evidence of construct validity for the C-Test. However, the C-Test failed to classify the subjects in their appropriate proficiency levels. A number of researchers (such as Dörnyei and Katona 1992; Klein-Braley 1997) have challenged the reliability of using C-Tests for placement purposes.

Mahfoud (2021) examined the face validity of the PT used at a Libyan HEI by using questionnaires and interviews. He also examined content validity by comparing PT and mid-term results. The findings revealed a high failure rate in the mid-term exam when the speaking and listening components were eliminated from the total score. As far as the Omani context is concerned, the only study published on PT evaluation was conducted by Al-Adawi and Al-Balushi (2016), who investigated the face validity of their institutional PT using teachers' and students' perceptions of the English PT at Colleges of Applied Sciences (CAS), Oman. They also compared students' PT scores against their mid-term scores. Both face and content validity of CAS English PT ranged from low to moderate levels. Nevertheless, comparing scores of the mid-term exam against PT scores might not be the best method to test the effectiveness, since both tests are designed with different purposes and comprise different content and format.

Considering the strengths and limitations of the studies mentioned above, this study assessed the validity and reliability of all four language skills tests of a computer-based online PT. Moreover, this study also benchmarked the in-house PT against the IELTS. The following section details the methods adopted to study the validity and reliability of the PT.

3. Research Questions

RQ$_1$. *How valid is the in-house PT instrument in placing the students at the appropriate level of the English Foundation Program?*

RQ$_2$. *How reliable is the in-house PT instrument in placing the students at the appropriate level of the English Foundation Program?*

4. Materials and Methods

4.1. Context of the Study

The study was conducted at one of the largest private HEIs in Oman with a student population of approximately 5000. The institution offers undergraduate (UG) and postgraduate (PG) level courses in the disciplines of Engineering, Business, and Computing. The medium of instruction in all these disciplines is English. To enter the undergraduate programs, students are required to have a minimum score of 5.5 on the IELTS, or a score of 525 and above for paper-based, or 196 and above for computer-based TOEFL, or a minimum score of 60% on the institutional in-house PT. Otherwise, to acquire the desired level of English language proficiency, the students are required to take the FP, which is a one-year preparatory course undertaken by students who lack the required English language, mathematics, IT, and study skills to specialize in a particular academic discipline. In this paper, only the English component of the PT is discussed.

4.2. A Brief on the in-House Placement Test Used at the HEI

The purpose of the IELTS patterned in-house PT of the HEI under study is to exempt or place all candidates in one of the three levels of the FP. The question papers (QP) and answer keys (AK) are prepared by the skills coordinators and are internally and externally moderated as per the institutional moderation policy using the online automated Content Management System (CMS). Multiple versions of QPs are configured on the institutional Moodle-based online learning platform for conducting the online PT.

The reading test consists of two texts with multiple-choice questions (MCQs). The writing test comprises two open-ended questions. The students type answers on the computer in the space allocated for this purpose. The first question is the description of a graph or a process in 150 words and the second one is a 250-word argumentative or persuasive essay. The listening test is adapted from the IELTS listening exam and consists of MCQs. The speaking test consists of a conversation between the teacher and the candidate and carries three sections. The first section focuses on self-introduction. The two other sections are focused on the discussion around a general topic. The time allotted for reading, writing, and listening components is 40 min each. However, the speaking test duration is 10 min. The writing component is assessed using a detailed marking rubric comprising the main areas, namely, grammatical structures, coherence, use of appropriate vocabulary, and content. The Table 1 below summarises the PT specifications.

The results of listening and reading tests are generated automatically. The speaking test is conducted in the presence of two examiners. A detailed marking rubric is used, which comprises the four main areas, namely, content, language accuracy, body language, and delivery.

The institutional eligibility criteria for being placed at different levels of FP are a score of 0–24 for Level 1, 25–44 for Level 2, and 45–59 for Level 3. These cut-off scores (grade threshold) were decided through a pilot study conducted by the institution, where a group of markers evaluated the answer scripts. These ranges were further verified by conducting annual benchmarking of PT with IELTS.

Table 1. Placement test specifications.

Skills	Test Components	Mode	Duration	Marks	Total Marks
Reading	Two reading texts	Computer-based Online	40 min	30	100
Writing	Two questions: (i) Description of a chart/graph or a process using the visual (word limit-150 words); (ii) Essay writing-argumentative/persuasive genre (word limit-250 words)		40 min	70	
Listening	MCQs (based on IELTS pattern)		40 min	50	100
Speaking	Oral (conversation) test consisting of 3 Sections (based on IELTS pattern)		10 min	50	

According to the institutional policy, a candidate must first attempt the reading and writing components of the PT. If the candidates score an average of 60% or above in reading and writing, they are eligible to take the listening and speaking tests. The candidates who score 60% or above in all four language skills are exempted from the English component of FP.

This study is based on the results of the test taken by the prospective students in the Fall semester of the academic year 2020–2021.

4.3. Validity and Reliability Study of the PT

In this study, three types of validity tests were used to establish the validity of the in-house PT (refer to Table 2).

Table 2. Placement test validation.

Validity			
Type of Validity	**Mode/Test/Method**	**No of Participants**	
Content validity	Automated workflow created on Content Management System (CMS)	2 Internal moderators	1 External moderator
		(1 Faculty and GFP Programme Manager)	(External expert)
Criterion-related Concurrent Validity	Paper-based IELTS exam	29 students	
Face validity	Online survey	57 students	

4.3.1. Content Validity

The content validity of the PT was established through internal and external moderation of question papers (QPs) and answer keys (AKs) conducted via an automated secure online process on the CMS (see Figure 1). Role-based access is given only to those faculty members who are involved in the QP setting and moderation process. Once the documents are uploaded to the system, the process starts, and automatic emails and reminders are sent to the respective personnel in the loop. Once the documents are uploaded by the skills coordinators, they can be accessed by the internal moderators who can write comments using the online assessment approval form. The files are then accessed by the skills coordinators to make the suggested changes. Once the changes are made, the documents are accessed by the external moderator who can review the files and give comments. Then, the skills coordinators make the final changes suggested by the external moderator. The internal moderators' team includes subject experts who are senior faculty members from the English language teaching team. The external expert is the subject specialist from a public university of international repute in Oman.

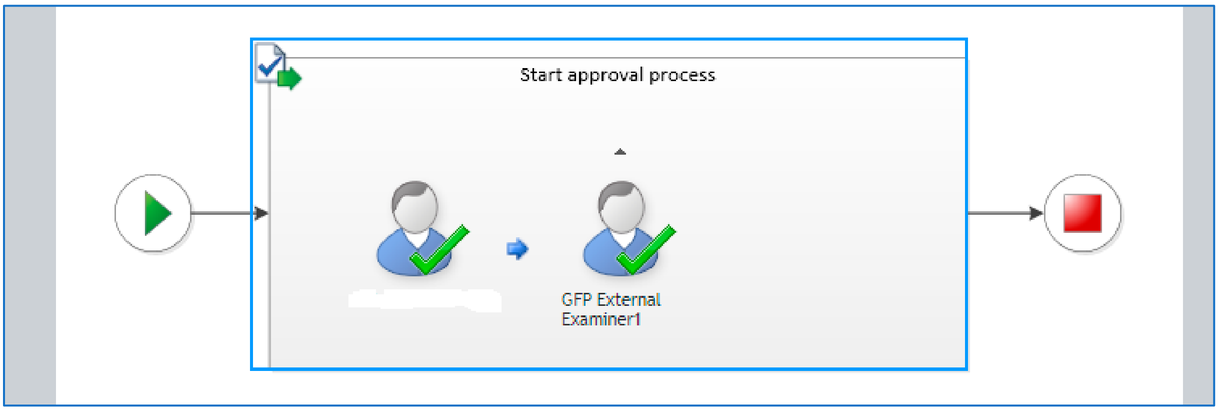

Figure 1. Snapshot of QP moderation workflow.

4.3.2. Criterion-Related Concurrent Validity

To ensure the effectiveness of the institutional English PT in terms of placing students appropriately into different levels of FP and guaranteeing entry and exit standards, as well as a progression between the three levels of FP, the English component of the PT is benchmarked against IELTS biannually. This exercise helps in measuring the criterion-related concurrent validity of the PT. The criterion-related validity of the Fall 2020 PT was established through an IELTS benchmarking exercise where a sample comprising 44 students was selected. The stratified sampling method was used; ten percent of the students from each of the three levels of the FP and ten percent of those who were directly placed in the Undergraduate programs were selected. However, out of 44, only 29 students appeared for the IELTS exam. It is important to note here that this test was conducted towards the end of the semester. The students were not given any formal training in IELTS before attempting the test.

4.3.3. Face Validity

To investigate the face validity of the test, student perceptions on the PT were gathered using a section of the institutional 'New Student Survey- Fall 2020' (MEC 2021). The section carried five items that used the Likert scale of agreement. The items included user-friendliness, process, the support provided during the test, duration, and overall satisfaction. The survey link was shared with all the students enrolled in the first semester of the Foundation, UG, and PG programs. However, only 109 students responded to the survey and 83 students answered the section on the PT. Only 57 (67.47%) students, who had attempted the PT, completed the PT section of the survey.

4.3.4. Reliability

For the reading test, Cronbach's alpha was used as a measure of internal consistency reliability of item interrelatedness to find out how closely related the items on the test are as a group and how well they contribute positively to measuring reading and comprehension ability. Marks of all the 306 students who completed the reading PT were analyzed. To establish the reliability of the writing assessment, 60 answer scripts, approximately 20% of the total responses to the writing test, were taken as samples for blind-marking by two examiners (refer to Table 3).

Table 3. Placement test reliability.

Skill	Method
Reading	Cronbach's alpha
Listening	Standard deviation
Writing	Paired sample *t*-test (dependent *t*-test) Double-blind marking
Speaking	Paired sample *t*-test (dependent *t*-test) conducted in the presence of two examiners

As mentioned earlier, the listening and speaking tests are taken by only those students who score an average of 60 and above on the reading and writing tests. This resulted in a small sample population which made it difficult to do the statistical calculation. It was not possible to use Cronbach's alpha for the listening test, since some component variables showed zero variance and were removed from the scale. Therefore, descriptive statistics, including mean and SD, were calculated using SPSS. Paired sample *t*-test (dependent *t*-test) was used to assess the inter-rater reliability of the speaking and writing tests.

5. Results

5.1. Criterion-Related Concurrent Validity: Benchmarking with Ielts

To confirm that the student placement into various levels of the FP (based on PT scores) and their exit from the FP is on par with the international standards, benchmarking with IELTS was conducted. A stratified sample of 29 students took the test towards the end of the Fall semester in 2020. To compare both results (PT marks and an IELTS band) precisely, the PT marks were converted to IELTS Bands using the following formula:

$$\text{Converted PT Mark} = \frac{\text{PT Mark}}{100} \times 9.$$

The results are presented in a cluster bar chart (Figure 2). It was noticed that the IELTS scores for most of the students are higher than their marks on PT (converted).

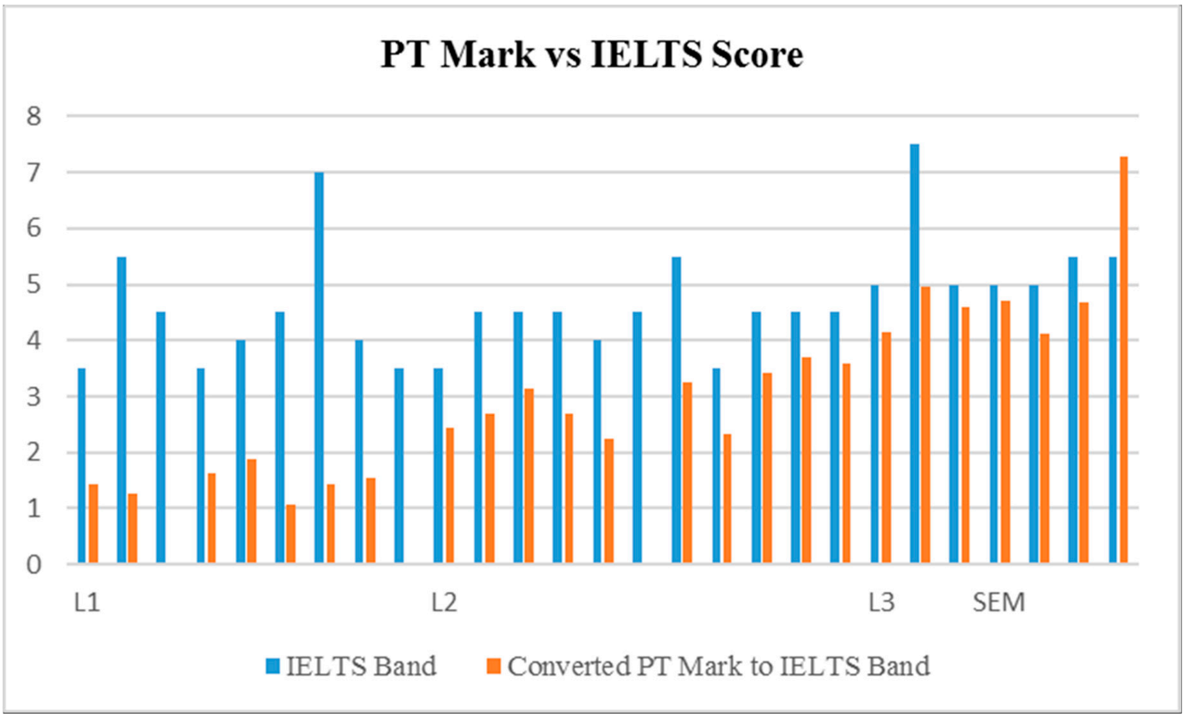

Figure 2. IELTS band and PT marks converted to IELTS band.

Regarding the Pearson correlation, as shown in Table 4 below, it was found that there is a positive moderate correlation between converted PT marks and the IELTS band which was statistically significant (r = 0.437, n = 27, p = 0.023). Due to the missing values, only 27 students' scores (out of 29) could be considered for calculating the Pearson correlation.

Table 4. Correlation between PT and IELTS.

Correlations			
		PT	Bands
PT	Pearson Correlation	1	0.437 *
	Sig. (2-tailed)		0.023
	N	27	27
Bands	Pearson Correlation	0.437 *	1
	Sig. (2-tailed)	0.023	
	N	27	27

Note: * Correlation is significant at the 0.05 level (2-tailed).

5.2. Content Validity: Internal and External Moderation

The test content was carefully reviewed by a panel of internal and external experts through an automated loop created on the CMS to determine content validity. The internal experts gave comments on test revision in terms of adding more sets for reading tests and the organization of content. This led to a few changes in the test before the link was shared with the external expert. The external expert, on the other hand, considered the appropriateness of tests for all four skills in terms of measuring the expected learning outcomes, the item difficulty, the language used, and the duration. The external expert also provided a post-moderation report with detailed comments which further strengthened the PT content.

5.3. Face Validity: Student Feedback on Placement Test

Figure 3 summarises the results of the PT section of the Institutional new student survey which was administered to the students enrolled in the first year of various foundation, UG, and PG programs. Out of the 88 students who had taken the in-house PT, 57 students attempted this section. The students were asked to rate the PT on a Likert scale of agreement where the responses ranged from strongly disagree to strongly agree (1–5). The results of all the items show a weighted average of 3.66 and above which established the face validity of the test.

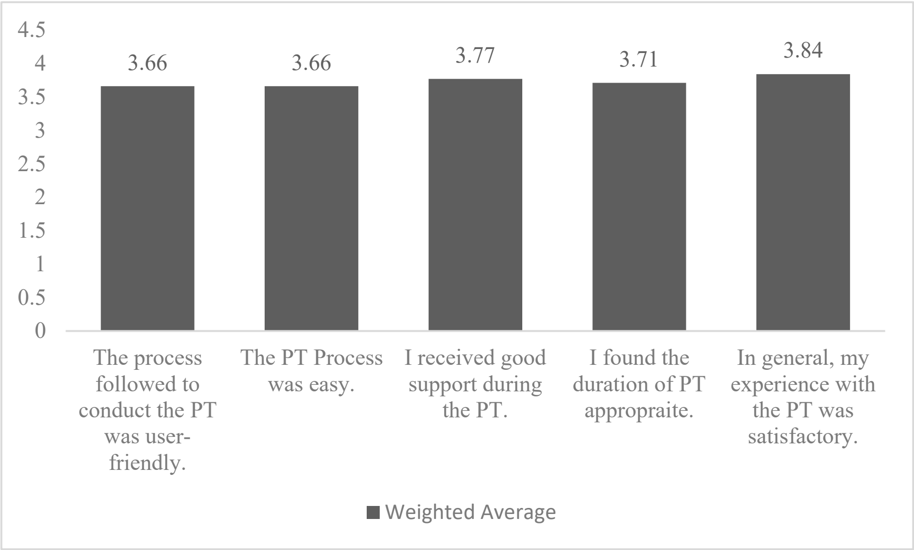

Figure 3. Student feedback on the placement test process.

5.4. Reliability and Validity of the Four English Language Components of the in-House PT

5.4.1. Reading Skills Test

The reliability of the reading skills test was calculated using Cronbach's Alpha. Cronbach's alpha (Cronbach 1951), also known as coefficient alpha, measures the reliability; more precisely, the internal consistency reliability or item interrelatedness, of a scale or test. Table 5 shows Cronbach's alpha values for the results obtained from six sets. As evident from the table, three sets show good reliability with Cronbach's alpha values above ≥0.8, for two sets it is acceptable (≥0.7), and for one set it appears to be questionable (≥0.6).

Table 5. Reading PT reliability: Cronbach's alpha.

Reading Placement Test	Cronbach's Alpha	Mean	Variance	Std. Deviation	N of Items
Set A	0.804 Good	17.12	51.172	7.153	17
Set B	0.804 Good	18.51	41.979	6.479	17
Set C	0.720 Acceptable	18.83	44.866	6.698	9
Set D	0.787 Acceptable	18.1364	47.609	6.89992	17
Set E	0.810 Good	21.8750	41.794	6.46480	17
Set F	0.646 Questionable	19.55	29.099	5.394	9

Note: "≥0.9—Excellent, ≥0.8—Good, ≥0.7—Acceptable, ≥0.6—Questionable, ≥0.5—Poor, and ≤0.5—Unacceptable" (George and Mallery 2003, p. 231).

5.4.2. Listening Skills Test

This section briefs the results obtained for the two sets of listening PT. Due to the nature of scores allotted for the listening test, the SD was found to be the most appropriate method to measure the reliability of the listening test. Tables 6 and 7 show the descriptive statistics including the SD for the two sets used. The number of students who undertook the 1st set was 15 and those who undertook the 2nd set was 33. The SD of the group of students who took Set A equals 2.2 and the SD of the group of students who took Set B equals 5.2. So, the SD of Set B is higher than Set A. In other words, the variation between the students' marks in Set B and the average mark is high as compared to students in Set A.

Table 6. Descriptive statistics—Set A.

	N	Minimum	Maximum	Mean	Std. Deviation
Total	15	41.25	48.75	46.0833	2.20929
Valid N (listwise)	15				

Table 7. Descriptive statistics—Set B.

	N	Minimum	Maximum	Mean	Std. Deviation
Total	33	28.75	48.75	44.5833	5.23535
Valid N (listwise)	33				

5.4.3. Writing Skills Test

To establish the reliability of the writing test, paired sample *t*-test (dependent *t*-Test) was used to compare the means of two related groups. The scores given by the first marker were compared with the scores given by the second marker after blind marking. As shown in Table 8 below, the significance value (*sig* = 0.163) is higher than the significance alpha (α = 0.05). Therefore, it can be safely concluded that there is no statistically significant difference between the first and the second marker.

Table 8. Paired samples test—writing.

		Paired Differences							
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		T	df	Sig. (2-Tailed)
					Lower	Upper			
Pair 1	Marker_1–Marker_2	0.850	4.664	0.602	−0.355	2.055	1.412	59	0.163

Note: alpha significance (α = 0.05).

5.4.4. Speaking Skills Test

The marks allotted by two examiners to all 68 students who took the speaking test were compared using paired sample t-tests (dependent t-test). It is clear from Table 9 below that the significance value (sig = 0.052) is greater than the significance alpha (α = 0.05). Hence, it can be concluded that there is no statistically significant difference between the marks allotted by the first and second examiners.

Table 9. Paired samples test—speaking.

		Paired Differences							
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		T	df	Sig. (2-Tailed)
					Lower	Upper			
Pair 1	Marker_1–Marker_2	−0.673	2.800	0.340	−1.351	0.005	1.981	67	0.052

Note: alpha significance (α = 0.05).

6. Discussion

The study set out to investigate the effectiveness of the in-house test in placing students at the appropriate levels of the FP by verifying its validity and reliability. To answer the first research question, three types of validity were considered, content, criterion-related, and face. Both internal and external moderators verified that the content is comprehensive enough and logically covers its intended variables. It is worth mentioning that the moderation workflow is configured in a way that the QP setter receives the QP after internal as well as external moderation. This allows the QP setter to incorporate changes based on the comments suggested by both internal and external moderators before the final version is configured on the learning management system (LMS). The final version, after the incorporation of all the suggestions, is considered to be a validated version. Therefore, there is a strong reason to believe that the moderation process of the PT ensures that the content is appropriate and relevant. From the judgments of the moderators, the measurable extent of each item for defining the traits and the set of questions that represent all aspects of the traits closely represented the PT requirements that are to be measured (Wallen and Fraenkel 2013). The results resonate with Nakamura's (2007) and Kim and Shin's (2006) studies that tested their models and technically verified the construct of the test.

Although the tests cannot be validated on content validity evidence solely, demonstration of content validity is a fundamental requirement of all assessment instruments. In this case, it is vital for the validation of PT and its conduction in the FP (Sireci 1998). It is worth mentioning, however, that the content validity of the PT might be questioned if MCQs are used as the predominant testing method. This is because MCQs might not provide a comprehensive picture of students' English proficiency levels. In addition, there are increasing chances of students cheating (Fan and Jin 2020). Fairness is another validity-related problem of the MCQs testing method (McCoubrie 2005; Xu et al. 2016). This could also be applicable to the listening and reading PT conducted at the research site although a number of QP sets are used. The writing PT, however, follows the IELTS pattern where no MCQs are given. Thus, the addition of the writing component to the PT gave more value to the test and improved its validity as it gave more focus on productive skills. In addition, the writing PT is a good method by which to align its content and methods with the teaching and learning objectives articulated in the curriculum (Fan and Jin 2020). Therefore, admittedly, the PT under study has a variety of testing methods that ensures a high validity of the test.

The data analysis of the student questionnaire revealed that face validity is satisfactory in terms of PT comprehensiveness, and relevance. The majority of participants were satisfied with the process of the conduction of the PT, with a weighted average between 3.66 to 3.84 on a scale of 5. Thus, the subjective assessment of PT provided a logical validity for measuring its cohesiveness and unambiguity. Yet, face validity is shallow and skims the surface only to form an opinion as it is criticized for being the weakest form of validity (e.g., Kaplan and Saccuzzo 2005; Weber 2004). For this reason, content validity was also conducted to establish a thorough examination of the PT's content, relevance, cohesiveness, and logical intent to measure the intended variable.

The criterion-related concurrent validity was verified by examining the relationship between PT and IELTS. It was found that the correlation is equal to r = 0.437, which means that the relationship is positively moderate. Moreover the participants' scores fell within the correct range of the IELTS bands which confirms the criterion validity coefficient as acceptable for most of the test items (Abdelkarim et al. 2021). Hence, it can be safely concluded that the concurrent validity has demonstrated favorable measures between the PT and the IELTS test (Fink 2010). This echoes Manganello's (2011) findings, which reported a moderate correlation coefficient between in-house English PTs and standardized tests such as TOFEL. Although the results of benchmarking with IELTS showed a positive moderate correlation, there are some disadvantages that lie within the criterion-based concurrent validity, including the fact that its estimates are likely to be smaller because of range restriction. Another potential concern regarding concurrent validation is the motivation of test-takers, a majority of whom are government-sponsored students, and many components of the PT, which include Maths and IT as well, are taken on the same day. The third area of concern is that the correlation coefficient between the IETS scores and PT might be affected depending on the interval gap between the two tests. That is, the wider the interval between the IELTS and the local PT, the smaller the correlation coefficient (Kokhan 2013). There is an interval of at least one semester (almost three months) between the PT and the IELTS. Therefore, this might have affected the correlation coefficient between the two scores.

Examining the second research question, the reliability test was conducted for all four skills including reading, writing, listening, and speaking. Cronbach's alpha was utilized for the reading PTs of six versions to examine the internal consistency in terms of how closely related the items of the test are. The results showed high coefficient figures and were good and acceptable for all the sets, except for one which is questionable (Shin and Lidster 2017), which might be due to the content of the reading text used. The figures for version A, B, and E were $\geq$0.8 which indicates that the coefficient of reliability is consistent between different versions. Versions C and D were also acceptable, with a figure of $\geq$0.7, which still indicates that there is constant reliability between different versions of the tests. Generally, most versions of the reading PT, except for one, have high internal consistency so the PT can be concluded as a reliable instrument to place students at their right level of English (Fan and Jin 2020).

Withe respect to listening test results, it was found that the SD of the group of students who took Set A, (2.2) is less than the SD of the group of students who took Set B (5.2). This means that the variation between students' marks from the mean in Set A of the listening test was lower, which could be due to the smaller sample size.

Regarding the reliability of writing and speaking, that was measured using paired sample *t*-tests, the results confirm that the comparison of the means of both the scores gave no significant statistical difference in values between the two markers, with a significance value of sig = 0.163 for the writing test and sig = 0.052 for the speaking test. The two raters were consistent in marking the productive skills, which is similar to the findings of Kim and Shin (2006). The statistical analysis compared the means of two markers in the writing and speaking test parametrically by defining the probability of variable distribution and making inferences about the parameters of the close distribution. Therefore, it can be concluded that, although the statistics of the *T*-test are about probabilities in the small difference between the two markers of the two PTs, which could be based on absolute standards, it is rather an evaluation of the probability of how close the markers were in their scores.

7. Limitations and Implications for Further Study

Since no tests are perfect in all aspects, the same can be applied here as well. The small sample size limited the generalizability of the findings. To further strengthen the reliability and validity of the test, a longitudinal study on a similar but larger sample population is required. This will enable a comparison of test results between semesters (Fulcher 1997). Validation could be conducted each year until enough QP sets have been developed to ensure accuracy. Similarly, more versions can be created and equated with validated ones. Since the listening and speaking tests are taken by only those students who score 60 and above in reading and writing and those who wish to be exempted from the FP, the sample size was rather small and, therefore, the findings cannot be generalized. Moreover, multiple versions of the test were created to maintain test security and facilitate administration. Since these versions are designed to be parallel, their strength can be further established via the linear equating method by comparing the SD and means of the scores attained on different versions (Buras 1996). Future validation exercises should consider teacher perceptions alongside student perceptions to ascertain face validity. Despite the said limitations, the study has strong implications for test developers in terms of the considerations in test design and evaluation. Moreover, the issues discussed, and procedures adopted here, provide educational policymakers with a comprehensive research-based approach to the systematic validation of PTs, which can be taken forward in formulating guidelines for validating PTs.

8. Conclusions

Placement testing approaches are dynamic and evolving and therefore require continual maintenance and validation. This study assessed the validity and reliability of an in-house online/computer-based PT. A combination of statistical methods was employed to establish the reliability, face validity, and criterion-related validity of the tests, while the content validity was established via internal and external moderation. Although the results of the PT are reliable overall, the small sample size limited the generalizability of the findings. However, there is a good reason to believe that the PT discussed here fulfills the purpose of placing students at different levels of FP to a large extent since the statistical analysis showed positive results for most of the test versions used for different skills. The strength of the test was further verified by internal and external moderation and IELTS benchmarking.

Author Contributions: All the authors of this paper have significantly contributed to the design, implementation, and compilation of this research. The research was conceptualized by S.N., R.S. and T.A.D.; and the methodology was devised by S.N. A.A.A. (Amal Al Amri) conducted the Formal analysis of the data. R.S. and S.A.A. administered the tests and conducted the IELTS benchmarking exercise. T.A.D. and A.A.A. (Asma Al Aufi) compiled the literature review section. The original draft of the manuscript was prepared by S.N. in collaboration with all the authors. The manuscript was reviewed by S.N. and T.A.D. S.N. is the Principal Investigator (PI) of the project that was selected for

funding. R.S. is the Co-PI, and T.A.D., S.A.A. and A.A.A. (Asma Al Aufi) are the project participants. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Research Council Oman, grant number BFP/RGP/EHR/20/439.

Institutional Review Board Statement: The study was conducted in accordance with the institutional policy on Research Ethics and Biosafety.

Informed Consent Statement: Informed consent was obtained from all the students who responded to the New Student Survey which was used to verify the face validity.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

Abdelkarim, Osama, Fritsch Julian, Jekauc Darko, and Bös Klaus. 2021. Examination of Construct Validity and Criterion-Related Validity of the German Motor Test in Egyptian Schoolchildren. *International Journal of Environmental Research and Public Health* 18: 8341. [CrossRef] [PubMed]

Al-Adawi, Sharifa Said Ali, and Aaisha Abdul Rahim Al-Balushi. 2016. Investigating Content and Face Validity of English Language Placement Test Designed by Colleges of Applied Science. *English Language Teaching (Online)* 9: 107–21. [CrossRef]

Ashraf, Hamid, and Samaneh Zolfaghari. 2018. EFL Teachers' Assessment Literacy and Their Reflective Teaching. *International Journal of Instruction* 11: 425–36. [CrossRef]

Buras, Avery. 1996. *Test Equating Procedures: A primer on the Logic and Applications of Test Equating*. ERIC Document Reproduction Service No. ED 395 038. New Orleans: Southwest Educational Research Association.

Chun, Jean Young. 2011. The Construct Validation of ELI Listening Placement Tests. *Education Psychology* 30: 1–47. Available online: http://www.hawaii.edu/sls/wp-content/uploads/2014/09/Chun-Jean-Young1.pdf (accessed on 21 October 2022).

Chung, Sun Joo, Haider Iftekhar, and Boyd Ryan. 2015. The English placement test at the University of Illinois at Urbana-Champaign. *Language Teaching* 48: 284–87. [CrossRef]

Coombe, Christine, Vafadar Hossein, and Mohebbi Hassan. 2020. Language assessment literacy: What do we need to learn, unlearn, and relearn? *Language Testing in Asia* 10: 1–16. [CrossRef]

Cronbach, Lee Joseph. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297–334. Available online: http://cda.psych.uiuc.edu/psychometrika_johnson/CronbachPaper%20(1).pdf (accessed on 20 September 2022). [CrossRef]

Dimova, Slobodanka, Yan Xun, and Ginther April. 2022. Local tests, local contexts. *Language Testing* 39: 341–54. [CrossRef]

Dong, Manxia, Cenyu Gan, Yaqiu Zheng, and Runsheng Yang. 2022. Research Trends and Development Patterns in Language Testing Over the Past Three Decades: A Bibliometric Study. *Frontiers in Psychology* 13: 1–15. [CrossRef]

Dörnyei, Zoltan, and Lucy Katona. 1992. Validation of the C-test amongst Hungarian EFL Learners. *Language Testing* 9: 187–206. [CrossRef]

Emperador-Garnace, Xenia Ribaya. 2021. Speaking Assessments in Multilingual English Language Teaching. *Online Submission* 25: 39–65. Available online: https://files.eric.ed.gov/fulltext/ED620449.pdf (accessed on 22 October 2022).

Fan, Jason, and Yan Jin. 2020. Standards for language assessment: Demystifying university-level English placement testing in China. *Asia Pacific Journal of Education* 40: 386–400. [CrossRef]

Fink, Arlene. 2010. Survey research methods. In *International Encyclopedia of Education*. Amsterdam: Elsevier, pp. 152–60. [CrossRef]

Fox, Jana D. 2009. Moderating top-down policy impact and supporting EAP curricular renewal: Exploring the potential of diagnostic assessment. *Journal of English for Academic Purposes* 8: 26–42. [CrossRef]

Fulcher, Glenn. 1997. An English Language placement test: Issues in reliability and validity. *Language Testing (Online)* 14: 113–38. Available online: http://languagetesting.info/articles/store/Placement%20Testing.pdf (accessed on 8 October 2022). [CrossRef]

Fulcher, Glenn. 1999. Computerizing an English language placement test. *ELT Journal* 53: 289–99. [CrossRef]

Fulcher, Glenn, Ali Panahi, and Hassan Mohebbi. 2022. Language Teaching Research Quarterly. *Language Teaching Research* 29: 20–56. [CrossRef]

Genç, Eda, Hacer Çalişkan, and Dogan Yuksel. 2020. Language Assessment Literacy Level of EFL Teachers: A Focus on Writing and Speaking Assessment. *Sakarya University Journal of Education* 10: 274–91. [CrossRef]

George, Darren, and Paul Mallery. 2003. *SPSS for Windows Step by Step: A Simple Guide and Reference 11.0 Update*, 4th ed. Boston: Allyn and Bacon.

Hilgers, Aimee. 2019. Placement Testing Instruments for Modality Streams in an English Language Program. Ph.D. thesis, Minnesota State University Moorhead, Moorhead, MN, USA.

Hille, Kathryn, and Yeonsuk Cho. 2020. Placement testing: One test, two tests, three tests? How many tests are sufficient? *Language Testing* 37: 453–71. [CrossRef]

Huang, Becky H., Mingxia Zhi, and Yangting Wang. 2020. Investigating the Validity of a University-Level ESL Speaking Placement Test via Mixed Methods Research. *International Journal of English Linguistics* 10: 1–15. [CrossRef]

Jamieson, Jeremy P., Matthew K. Nock, and Wendy Berry Mendes. 2013. Improving acute stress responses: The power of reappraisal. *Current Directions in Psychological Science* 22: 51–56. [CrossRef]

Johnson, Robert C., and A. Mehdi Riazi. 2017. Validation of a Locally Created and Rated Writing Test Used for Placement in a Higher Education EFL Program. *Assessing Writing* 32: 85–104. [CrossRef]

Kane, Michael T. 2013. Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement* 50: 1–73. [CrossRef]

Kaplan, Robert Malcolm, and Dennis P. Saccuzzo. 2005. *Psychological testing: Principles, Applications and Issues*, 6th ed. Belmont: Thomson Wadsworth.

Kim, Young-Mi, and Misook Kim. 2017. Validations of an English Placement Test for a General English Language Program at the Tertiary Level. *JLTA Journal* 20: 17–34. [CrossRef]

Kim, Hyun Jung, and Hye Won Shin. 2006. A reading and writing placement test: Design, evaluation, and analysis. *Studies in Applied Linguistics and TESOL* 6: 2.

Klein-Braley, Christine. 1997. C-tests in the context of reduced redundancy testing: An appraisal. *Language Testing* 14: 47–84. [CrossRef]

Kokhan, Kateryna. 2013. An argument against using standardized test scores for placement of international undergraduate students in English as a Second Language (ESL) courses. *Language Testing* 30: 467–89. [CrossRef]

Li, Zhi. 2015. Using an English self-assessment tool to validate an English Placement Test. *Language Testing and Assessment* 4: 59–96. Available online: https://arts.unimelb.edu.au/__data/assets/pdf_file/0003/1770672/Li.pdf (accessed on 1 October 2022).

Liao, Yen-Fen. 2022. Using the English GSAT for placement into EFL classes: Accuracy and validity concerns. *Language Testing in Asia* 12: 1–23. [CrossRef]

Liskinasih, Ayu, and Rizky Lutviana. 2016. The validity evidence of TOEFL test as placement test. *Jurnal Ilmiah Bahan dan Sastra* 3: 173–80. Available online: https://www.researchgate.net/publication/314110391 (accessed on 25 October 2022). [CrossRef]

Mahfoud, Bashir Ghit. 2021. Examining the Content and Face Validity of English Placement Test at the Technical College of Civil Aviation and Meteorology, Tripoli, Libya. *AL-JAMEAI* 33: 5–19. Available online: https://www.aljameai.org.ly/index.php/aljameai/article/view/802 (accessed on 11 September 2022).

Manganello, Marc. 2011. Correlations in the New Toefl Era: An Investigation of the Statistical Relationships Between Ibt Scores, Placement Test Performance, and Academic Success of International Students at Iowa State University. Unpublished Master's thesis, Iowa State University, Ames, IA, USA.

McCoubrie, Paul. 2005. Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher* 26: 709–12. [CrossRef] [PubMed]

MEC. 2021. *Unpublished Raw Data on New Student Survey*. Seeb: Middle East College.

Messick, Samuel. 1996. Validity and washback in language testing. *Language Testing* 13: 241–56. Available online: https://files.eric.ed.gov/fulltext/ED403277.pdf (accessed on 15 September 2022). [CrossRef]

Nakamura, Yuji. 2007. A Rasch-based analysis of an in-house English placement test. In Paper presented at the Second Language Acquisition—Theory and pedagogy: Proceedings of the 6th annual JALT Pan-SIG Conference, Online, May 12–13; pp. 97–109.

Rahal, Hadeel El, and Huda Dimashkie. 2020. Creational Reverse Engineering: A Project to Enhance English Placement Test Security, Validity, and Reliability. In *The Assessment of L2 Written English across the MENA Region*. London: Palgrave Macmillan, pp. 43–68.

Rouhani, Mahmood. 2008. Another look at the C-Test: A validation study with Iranian EFL learners. *The Asian EFL Journal Quarterly March* 10: 154.

Shin, Sun-Young, and Ryan Lidster. 2017. Evaluating different standard-setting methods in an ESL placement testing context. *Language Testing* 34: 357–81. [CrossRef]

Sireci, Stephen. G. 1998. The construct of content validity. *Social Indicators Research* 45: 83–117. Available online: https://www.jstor.org/stable/27522338 (accessed on 15 September 2022). [CrossRef]

Topor, F. Sigmond. 2014. A sentence repetition placement test for ESL/EFL learners in Japan. In *Handbook of Research on Education and Technology in a Changing Society*. Hershey: IGI Global, pp. 971–988. [CrossRef]

Wall, Dianne, Caroline Clapham, and J. Charles Alderson. 1994. Evaluating a placement test. *Language Testing* 11: 321–44. [CrossRef]

Wallen, Norman E., and Jack R. Fraenkel. 2013. *Educational Research: A guide to the Process*. Beijing: Routledge.

Weber, Robert Philip. 2004. Content analysis. In *Social Research Methods: A Reader*. Edited by Clive Seale. Oxford: Routledge, pp. 117–24.

Westrick, Paul. 2005. Score Reliability and Placement Testing. *JALT Journal* 27: 71–94. Available online: https://jalt-publications.org/sites/default/files/pdf-article/jj-27.1-art4.pdf (accessed on 20 October 2022). [CrossRef]

Xu, Xiaomeng, Sierra Kauer, and Samantha Tupy. 2016. Multiple-choice questions: Tips for optimizing assessment in-seat and online. *Scholarship of Teaching and Learning in Psychology* 2: 147. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.