Reliability and Validity of Online Placement Test

Subjects: Linguistics

Contributor: Samia Naqvi, Reema Srivastava, Tareq Al Damen, Asma Al Aufi, Amal Al Amri, Suleiman Al Adawi

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. 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.

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

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).

2. Reliability and Validity of Online Placement Test

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; Genc 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.

References

1. 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.
2. 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).
3. Fulcher, Glenn, Ali Panahi, and Hassan Mohebbi. 2022. Language Teaching Research Quarterly. Language Teaching Research 29: 20–56.

4. Hille, Kathryn, and Yeonsuk Cho. 2020. Placement testing: One test, two tests, three tests? How many tests are sufficient? Language Testing 37: 453–71.
5. Liao, Yen-Fen. 2022. Using the English GSAT for placement into EFL classes: Accuracy and validity concerns. Language Testing in Asia 12: 1–23.
6. Shin, Sun-Young, and Ryan Lidster. 2017. Evaluating different standard-setting methods in an ESL placement testing context. Language Testing 34: 357–81.
7. 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.
8. 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.
9. 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.
10. 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.
11. Wall, Dianne, Caroline Clapham, and J. Charles Alderson. 1994. Evaluating a placement test. Language Testing 11: 321–44.
12. Hilgers, Aimee. 2019. Placement Testing Instruments for Modality Streams in an English Language Program. Ph.D. thesis, Minnesota State University Moorhead, Moorhead, MN, USA.
13. 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).
14. Dimova, Slobodanka, Yan Xun, and Ginther April. 2022. Local tests, local contexts. Language Testing 39: 341–54.
15. 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.
16. 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.
17. 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).
18. 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).
19. Fulcher, Glenn. 1999. Computerizing an English language placement test. ELT Journal 53: 289–99.
20. 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.
21. 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.
22. 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.
23. 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).
24. Kane, Michael T. 2013. Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement 50: 1–73.
25. 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.
26. Ashraf, Hamid, and Samaneh Zolfaghari. 2018. EFL Teachers' Assessment Literacy and Their Reflective Teaching. International Journal of Instruction 11: 425–36.
27. 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.
28. 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.

29. 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).
30. 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).
31. 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.
32. 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.
33. 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).
34. 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.
35. Rouhani, Mahmood. 2008. Another look at the C-Test: A validation study with Iranian EFL learners. The Asian EFL Journal Quarterly March 10: 154.
36. Dörnyei, Zoltan, and Lucy Katona. 1992. Validation of the C-test amongst Hungarian EFL Learners. Language Testing 9: 187–206.
37. Klein-Braley, Christine. 1997. C-tests in the context of reduced redundancy testing: An appraisal. Language Testing 14: 47–84.