Fonologi: Jurnal Ilmuan Bahasa dan Sastra Inggris
Vol. 2, No. 3 September 2024
e-ISSN: 3025-6003, p-ISSN: 3025-5996, Hal 325-332
DOI: https://doi.org/10.61132/fonologi.v2i2.1107
Available online at: https://journal.aspirasi.or.id/index.php/Fonologi

Examining Validity and Reliability of General English Course Assessments in Higher Education

Abila Meika Kurnia Putri[1*], Barlian Kristanto[2], Muhammad Soali[3]
[123] Universitas Harapan Bangsa, Indonesia
[1*]abilameikaa@gmail.com, [2]barliankristanto@gmail.com, [3]muhammadsoali@uhb.ac.id

Address: JL. K.H. Wahid Hasyim, No. 274-A, Windusara, Karangklesem, Purwokerto Sel. District, Banyumas Regency, Central Java 53144
Author correspondence: abilameikaa@gmail.com

Abstract. *This study investigated the validity and reliability of a General English Course assessment tool at Universitas Harapan Bangsa. Using a quantitative descriptive technique, the study examined the assessment items using point-biserial correlation for validity and Cronbach's alpha for reliability. While the instrument had high overall reliability, several items needed to be revised to improve its validity. The findings indicate that the assessment instrument is generally reliable and valid; nevertheless, continuing monitoring and adjustment are required to maintain its quality and assure accurate evaluation of student English competence.*

Keywords: *Language assessment, validity, reliability, EFL (English as a foreign language)*

1. INTRODUCTION

Language assessment is important in higher education, especially when assess English competency in the context of English as a Foreign Language (EFL). The validity and reliability of these assessment are critical to ensure a fair and accurate appraisal of students' abilities. Validity refers to the correctness of what a test measures, whereas reliability refers to the consistency and repeatability of test results (Weideman, 2019) .These two factors have been the key emphasis in the creation of conceptual frameworks for language assessment (Mohajan, 2017).

The conceptual framework for validity in language assessment has changed greatly, reflecting a variety of techniques and procedures. This approach is critical for ensuring that language tests effectively assess their intended purpose and are utilized responsibly in educational settings (Weideman, 2019). According to Bennett & Wood (1987), this theory combines content, criterion-related, and construct validity to provide a holistic validity idea. The application of this idea of validity into practice, this study will use point-biserial correlation, a statistical approach that assesses the link between performance on individual test items and overall test score. This approach is quite successful in determining the amount to which each test question contributes to measuring the desired construct, providing valuable insights into the assessment instrument's validity (Lin & Chang, 2019).

Received: Agustus 30, 2024; Revised: September 14, 2024; Accepted: September 28, 2024; Published: September 30, 2024

Whilst the validity of a test is paramount, reliability is highly important when it comes to language assessment (Tosuncuoglu, 2018). Based on Brown (2010), reliability assessment usually include internal consistency, test-retest reliability and inter-rater reliability. Reliability will be considered in the measurement of Cronbach's alpha which is commonly used to check for internal consistency evaluation tools. Cronbach's alpha is a coefficient that quantifies how reliable the items of a test are an index that provides crucial information about the overall reliability of the measuring instrument (Arikunto, 2013). This use of single-factor model is consistent with the current best research practices in contemporary language assessment.

Recent research in EFL contexts has highlighted the relevance of assessing validity and reliability in language evaluations. According to Moafian, et al (2018), emphasized the need of a validation framework that considers both theoretical and practical elements of language testing. Mohajan (2017) and Lin & Chang (2019), conducted research on the accuracy and reliability of English language tests in higher education, offering useful insights into the field's difficulties and prospects. These findings have obvious consequences for higher education institutions that provide English language programs.

General English Course at Universitas Harapan Bangsa is a course of the university's Centre for Language Development (P2B) programme. The course is designed to prepare students for academic and professional success on the global stage with advanced English language proficiency required by all academic departments. This is particularly timely, given that the field of education and employment on a worldwide communication scale and in global academia continues to grow at a rapid pace and English language proficiency is becoming a fundamental necessity. Not surprisingly, with English being such a dominant global language as a lingua franca, we must be sure that our assessments are measuring something very important. Assessment errors can lead to huge differences in students' academic and career paths; not to mention the competitiveness of a country.

Recognizing the significance of the assessment tool's quality, which has an influence on a variety of factors, including examinee quality. The research intends to give insight into the quality of the evaluation instrument employed at Universitas Harapan Bangsa by examining the questions' validity and reliability.

2. THEORETICAL REVIEW

Validity

ValidityA test is deemed to have high validity when the instrument performs its measuring function or produces accurate results that correlate to what is being assessed (Arifin, 2019). The validity test determines the amount of accuracy and attention necessary to verify that an assessment instrument performs its measuring function. The validity of the questions is determined by comparing each question gained to the overall score.

The point correlation calculation results are then compared to r_{table} at 5% significance threshold, adjusted for number percentages. If $Y_{pbi} > r_{table}$, the items can be considered accurate. The questions can be said valid. As well as vice versa. The point biserial correlation technique (Ypbi) can be utilized for validity testing (Arikunto, 2013). The point-biserial correlation equation is represented as follows:

$$Y_{pbi} = \frac{M_p - M_t}{S_t}\sqrt{\frac{p}{q}}$$

Description:

Y_{pbi} = Point biserial correlation.

M_p = The average score of participants correctly answering the item under validity review.

M_t = Total score.

S_t = The standard deviation of the entire score.

p = The proportion of learners who answered correctly.

q = The fraction of learners who answered wrongly.

Reliability

According to Arifin (2019) points out that reliability in testing is an important indicator of a test instrument's consistency. A dependable tool gives consistent findings when delivered to the same group at different periods. A numerical coefficient is used to determine dependability, with larger values indicating more reliability. Arifin defines a successful reliability test as having a high coefficient and a low standard error of measurement. The Alpha Cronbach approach is expressed by the formula:

$$a = \frac{R}{R-1}\left(1 - \frac{\sum {\sigma^2}_i}{{\sigma_x}^2}\right)$$

Description:

α = Alpha coefficient

R = number of items

σ^2_i = items variance

σ_x^2 = total score variance

According to Sudijono, as described by Pamungkasih & Nawawi (2021), an Alpha coefficient (α) of 0.70 or greater indicates highly trustworthy test items, whereas a coefficient less than 0.70 shows unreliability. This standard assists researchers and educators in ensuring that their assessment tools consistently evaluate desired outcomes, resulting in more accurate and reliable results across a wide range of academic disciplines.

3. RESEARCH METHODOLOGY

The researchers used a descriptive quantitative method to assess General English Course evaluations for beginner anesthesiology students at Harpan Bangsa University in 2023/2024. The study employed cluster sampling and collected data through documentation of existing assessment materials. Analysis involved spreadsheet software for initial organization and SPSS version 23 for advanced statistical techniques, including item analysis, factor analysis for validity, and Cronbach's alpha for reliability. This comprehensive approach aimed to objectively evaluate the assessments, identifying strengths and areas for improvement in language assessment practices within this academic setting.

4. RESULT AND DISCUSSION

Result

Validity

The item validity analysis conducted on this measuring instrument provides the following results, as shown in the tables 1:

Table 1 Distribution of test item based on validity

Item Category	Number of Items	Percentage
Accepted	34	85%
Need Revision	2	5%
Poor	4	10%
Total	40	100%

The results showed that 34 of the 40 multiple-choice questions examined (85%) were assessed as accepted, indicating acceptable construct validity and acceptability for application. However, two items (5%) were identified as requiring improvements, suggesting that their construct validity may be improved. Furthermore, four items (10%) were classified as unacceptable, suggesting a lack of construct validity and needing major adjustment or elimination from the exam.

Reliability

The research showed a high level of overall reliability. With a Cronbach's Alpha rating of 0.86, the test had strong internal consistency, indicating that the questions accurately measured the target construct of beginner-level English competence. The Item-Total Statistics showed that the majority of questions contributed positively to the scale's reliability, with adjusted item-total correlations ranging from 0.004 to 0.576. Table 2 shows how the reliability index helped to categorize the distribution of test items.

Table 2 Distribution of test item based on reliability

Item	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted	Impact on Scale Reliability
X1-X2, X5-X7, X9-X12, X15-X40	Various (0.227 to 0.576)	0.851 to 0.859	Positive contribution
X3	0.125	0.860	Slight negative impact
X4	0.069	0.863	Negative impact
X8	0.004	0.865	Negative impact
X13	0.241	0.859	Neutral
X14	0.230	0.859	Neutral

Discussion

Validity

The item validity analysis of the assessment instrument produced positive findings, with 34 items (85%) categorized as excellent. This is consistent with Arifin's (2019) criterion for high validity, in which an instrument properly assesses its target concept. However, six elements need attention: two (5%) required improvement, and four (10%) were ruled

inadequate. Arikunto's (2013) point biserial correlation approach found that these items had a low correlation with the total score, indicating the need for modifications. Comparative research revealed somewhat higher validity rates: Alaofi & Russell (2022), identified 90% relevance in a computer terminology test, while Lin and Chang (2019) discovered 93% validity in an English subtest. These benchmarks point to areas where the existing instrument might be improved.

According to Louch et al. (2019), suggestion to replace or delete items with correlation values less than 0.30, and Arhin et al. (2023), emphasis on frequent item analyses, the instrument needs some refining to improve its overall validity. The validity ratings (score 0: 3 items, score 1: 2 items, score 2: 34 items) point to adopting the alternative hypothesis (H1) and rejecting the null hypothesis (H0), demonstrating excellent overall validity. However, in order to reach the validity standards of comparable research, problematic items must be revised and further analysis performed to improve the instrument's accuracy and consistency with its measuring aims.

Reliability

The reliability analysis of the 40-item multiple-choice test revealed great internal consistency, with a Cronbach's Alpha of 0.86, higher above the required criterion of 0.70 for research purposes. This high reliability coefficient suggests that the Harapan Bangsa University assessment regularly reflects beginner-level English competence, which aligns with course goals and serves as a solid standard for student evaluation. This reliability is comparable to or somewhat better than similar studies in the area, such as Trivict & Densiana (2020) research, which reported a coefficient of 0.820, and Arini & Dzulfikri's (2022) study, which reported 0.91.

The test's high reliability has positive effect on university decision-making, allowing educators to confidently make informed assessments of students' language competency, placement, and course completion. It provides the principles for fair and equal treatment of pupils by minimizing the impact of measurement inaccuracy. However, items X3, X4, and X8 performed poorly with low item-total correlations, and items X13 and X14 had a neutral influence, indicating places for development.

Overall, with 37 out of 40 items showing reliable, the analysis suggests accepting the alternative hypothesis (H1) and rejecting the null hypothesis (H0), so establishing the scale's

dependability. To improve the test's already high reliability, update or remove items X3, X4, and X8, as well as review items X13 and X14, as indicated by Brown's (2010) scale improvement suggestions.

5. CONCLUSION AND SUGGESTION

Conclusion

Validity

The assessment instrument demonstrated high overall validity, with 85% of items classified as good. However, there was room for improvement, as 5% of items required adjustment and 10% were evaluated as poor. The analysis supported accepting the alternative hypothesis (H1), indicating that the instrument had strong validity. Nevertheless, to enhance its accuracy and alignment with measurement goals, problematic items needed to be revised. This process would further strengthen the instrument's ability to consistently assess the intended construct.

Reliability

The instrument showed strong reliability, evidenced by a Cronbach's Alpha of 0.86, which exceeded the recommended threshold of 0.70. With 37 out of 40 items multiple-choice proving reliable, the analysis supported accepting the alternative hypothesis (H1), confirming the scale's overall reliability. However, items X3, X4, and X8 performed poorly and needed to be considered for revision or removal. By addressing these weaker items, the already high reliability of the scale could be further improved, ensuring more consistent and dependable measurement results across various applications of the instrument.

Suggestion

To improve the general validity and reliability of the assessment instrument, test developers should modify the poor items, notably X3, X4, and X8. To maintain consistent instrument quality, test developers should use a comprehensive and continuous approach. This entails assessing cognitive abilities, consulting specialists, doing pilot tests, implementing continual improvement, and keeping extensive records. By regularly adhering to these suggestions, the instrument's quality and efficacy will increase dramatically, resulting in a more accurate and meaningful evaluation of student performance.

REFERENCES

Alaofi, S., & Russell, S. (2022). A Validated Computer Terminology Test for Predicting Non-native English-speaking CS1 Students' Academic Performance. *ACM International Conference Proceeding Series*, 133–142. https://doi.org/10.1145/3511861.3511876

Arhin, D., Annan-Brew, R., & Owusuaah, R. (2023). Exploring the Quality of Multiple-Choice Question Type of Test Items in Information and Communication Technology Using Item Analysis. *E-Journal of Humanities, Arts and Social Sciences*, *4*(1), 50–58. https://doi.org/10.38159/ehass.2023414

Arifin, Z. (2019). Evaluasi Pembelajaran. In *Bandung. PT Remaja Rosdakarya*.

Arikunto, S. (2013). Dasar-Dasar Evaluasi Pendidikan. *Jakarta : Bumi Aksara*.

Arini, M. A. D., & Dzulfikri. (2022). Interrogating the Quality of Junior High School Final Exam Made by EFL Teacher Using ITEMAN 4.3 Program. *EDUTEC : Journal of Education And Technology*, *6*(2), 294–306. https://doi.org/10.29062/edu.v6i2.425

Bennett, M., & Wood, R. (1987). Measurement and Assessment in Education and Psychology. *European Journal of Education*, *22*(3/4), 363. https://doi.org/10.2307/1502911

Brown, H. D. & A. P. (2010). *Language Assessment Principle and Classroom Practice* (Third Edit). White Plains,NY: Pearson Education, Inc.

Lin, W. Y., & Chang, Y. J. (2019). Construct validation of the multiple-choice items of the English subtest of the advanced subjects test in Taiwan. *Electronic Journal of Foreign Language Teaching*, *16*(1), 80–94.

Louch, G., Reynolds, C., Moore, S., Marsh, C., Heyhoe, J., Albutt, A., & Lawton, R. (2019). Validation of revised patient measures of safety: PMOS-30 and PMOS-10. *BMJ Open*, *9*(11), 1–11. https://doi.org/10.1136/bmjopen-2019-031355

Moafian , F ., Ostovar , S ., Griffiths , M . D . & Hashemi, M. (2018). *The Construct Validity and Reliability of the Characteristics of Successful EFL Teachers Questionnaire (CoSEFLT-Q). Porta*. 1–13.

Mohajan, H. K. (2017). Two Criteria for Good Measurements in Research: Validity and Reliability. *Annals of Spiru Haret University. Economic Series*, *17*(4), 59–82. https://doi.org/10.26458/1746

Pamungkasih, R. S. N., & Nawawi, E. (2021). Analisis Kualitas Butir Soal Ujian Akhir Semester Ganjil pada Mata Pelajaran Kimia Kelas X di SMA Negeri 8 Palembang Tahun Ajaran 2020/2021. *Prosiding Seminar Nasional Pendidikan IPA Tahun 2021*, *1*.

Tosuncuoglu, I. (2018). Importance of Assessment in ELT. *Journal of Education and Training Studies*, *6*(9), 163. https://doi.org/10.11114/jets.v6i9.3443

Trivict, T., & Densiana, F. (2020). *The quality of an English summative test of a public junior*. *3*(2), 133–141.

Weideman, A. (2019). Degrees of adequacy: The disclosure of levels of validity in language assessment. *Koers*, *84*(1), 1–15. https://doi.org/10.19108/KOERS.84.1.2451