Lisanudhad: Jurnal Bahasa, Pembelajaran dan Sastra Arab
Vol. 11, No. 2 (2024), pp. 164-192, DOI: https://doi.org/10.21111/lisanudhad.v11i2.12163
P-ISSN: 2356-1998; E-ISSN: 2527-6360
https://ejournal.unida.gontor.ac.id/index.php/lisanu/index

Evaluation of the Feasibility and Reliability of Arabic Multiple Choice Tests in Higher Education

Zakiyah [a,1,*], Sulfiatin [b,2], M. Baihaqi [c,3], Nais Musyafaul Faiza[d,4], Muhammad Ilham Revo Alghani [e,5], Fatihuddin [f,6]

[a) b) c) d) e)] UIN Sunan Ampel Surabaya, [f)] Al Qasimy University Kingdom Saudi Arabia
[1]02040923020@uinsa.ac.id, [2]sulfiatitin@gmail.com, [3]baihaqi@uinsa.ac.id,
[4]02040923030@uinsa.ac.id, [5]revoilham71@gmail.com, [6]422117642@qu.edu.sa

Abstract

Multiple-choice tests are commonly used in education, including at the higher education level, as an efficient method for evaluating students' understanding. In the context of Arabic language learning, multiple-choice tests are trusted to be able to assess students' level of Language comprehension and mastery. However, it is essential to evaluate the feasibility and reliability of these tests to ensure that the results accurately reflect students' abilities. This study aims to evaluate the feasibility and reliability of multiple-choice tests in Arabic in higher education. The evaluation was conducted by assessing content validity, construct validity, reliability, and correlation with students' academic performance. The results showed that 60% of the test questions were aligned with the existing curriculum, although only 14 out of 25 questions met the criteria for construct validity. Despite some shortcomings in construct validity, the test demonstrated a high level of reliability, with a Cronbach's Alpha value of 0.88, indicating consistent test results. Additionally, there was a significant positive correlation between test scores and students' academic performance (r = 0.44), indicating that the test can reflect students' overall academic achievement. Despite certain limitations in construct validity, the conclusion of this study is that the multiple-choice test is still considered reliable as an evaluation tool. This conclusion provides insight into the test's effectiveness in measuring students' understanding and mastery of Arabic at the tertiary level.

Keywords: *Multiple Choice Test, Feasibility and Reliability Evaluation, Higher Education.*

تقييم جدوى وثبات اختبارات الاختيار من متعدد في اللغة العربية في التعليم العالي

زكية [أ-١*]، **سلفية**[ب-٢]، **محمد بهقي**[ج-٣]، **نايس مشفع الفائزة**[د-٤]، **محمد ريفو إلهام الغاني**[هـ-٥]، **فتيح الدين**[و-٦]

[أ ب ج د،هـ] جامعة الإسلامية الحكومية سونان أمبيل سورابايا، [و]جامعة القاسمي المملكة العربية السعودية

[١] sulfiatitin@gmail.com[٢]،02040923020@uinsa.ac.id

baihaqi@uinsa.ac.id[٣]،02040923030@uinsa.ac.id[٤]، revoilham71@gmail.com[٥]،
422117642@qu.edu.sa[٦]

الملخص

تُستخدم الاختبارات متعددة الخيارات بشكل شائع في التعليم، بما في ذلك في مستوى التعليم العالي، كوسيلة فعالة لتقييم فهم الطلاب. وفي سياق تعلم اللغة العربية، يُعتقد أن الاختبارات متعددة الخيارات قادرة على تقييم مستوى فهم وإتقان الطلاب للغة. ومع ذلك، من الضروري تقييم مدى ملاءمة وموثوقية هذه الاختبارات لضمان أن النتائج تعكس بدقة قدرات الطلاب. يهدف هذا البحث إلى تقييم مدى ملاءمة وموثوقية الاختبارات متعددة الخيارات في اللغة العربية في التعليم العالي. تم إجراء التقييم من خلال فحص صلاحية المحتوى، وصلاحية البنية، والموثوقية، والعلاقة مع الأداء الأكاديمي للطلاب. أظهرت النتائج أن ٦٠٪ من أسئلة الاختبار كانت متوافقة مع المنهاج الحالي، رغم أن ١٤ سؤالاً فقط من أصل ٢٥ استوفت معايير صلاحية البنية. وعلى الرغم من بعض القصور في صلاحية البنية، أظهر الاختبار مستوى عالياً من الموثوقية بقيمة ألفا كرونباخ بلغت ٠,٨٨، مما يشير إلى اتساق النتائج. بالإضافة إلى ذلك، كانت هناك علاقة إيجابية ذات دلالة إحصائية بين درجات الاختبار والأداء الأكاديمي للطلاب ($r = 0.44$)، مما يشير إلى أن الاختبار يمكن أن يعكس إنجاز الطلاب الأكاديمي العام. وعلى الرغم من بعض القيود في صلاحية البنية، فإن الاستنتاج النهائي لهذه الدراسة هو أن الاختبار متعدد الخيارات يُعتبر أداة تقييم موثوقة. يوفر هذا الاستنتاج رؤية حول فعالية الاختبار في قياس فهم وإتقان الطلاب للغة العربية على مستوى التعليم العالي.

الكلمات الرئيسية: اختبار الاختيار من متعدد، تقييم الجدوى والموثوقية، التعليم العالي

Introduction

In the scope of higher education, evaluation is one of the important aspects in testing the effectiveness of learning in educational purposes. A good evaluation is not only provides an explanation or description of students' understanding of the material taught, but also helps teachers or lecturers in assessing and measuring the level of success of students following the learning program that has been designed in higher education. In this case, Arabic language learning is selected with the suitable and accurate evaluation method of learner competence.[1]

One of the evaluation methods used in general is Multiple choice test. This test is frequently utilized both in school and university because of its ability to provide output data that can be assessed quantitatively. In this case, multiple choice tests are very efficient as an assessment of learning materials and students' understanding.[2] However, to ensure the validity and reliability of multiple choice tests in Arabic language learning in higher

[1] Maryam Safdari and Jalil Fathi, "Investigating the Role of Dynamic Assessment on Speaking Accuracy and Fluency of Pre-Intermediate EFL Learners," ed. Richard Kruk, *Cogent Education* 7, no. 1 (January 1, 2020): 1818924, https://doi.org/10.1080/2331186X.2020.1818924.

[2] Jalil Fathi, Lawrence Jun Zhang, and Mohammad Hossein Arefian, "Testing a Model of EFL Teachers' Work Engagement: The Roles of Teachers' Professional Identity, L2 Grit, and Foreign Language Teaching Enjoyment," *International Review of Applied Linguistics in Language Teaching* 0, no. 0 (July 21, 2023), https://doi.org/10.1515/iral-2023-0024.

education, an in-depth evaluation needs to be conducted to determine their feasibility and reliability.[3]

Research related to evaluating the feasibility and reliability of Arabic multiple choice tests in higher education has widespread and urgent. It is important to ensure that the tests used are capable of conducting evaluations that illustrate accuracy in students' understanding of Arabic language learning and make a basis for making decisions that interrelate learning and curriculum development. Therefore, this study aims to provide insight in evaluating multiple choice test methods in Arabic language learning in higher education.[4]

In fact, it is not all test tool suitable for test taker.[5] While tests are often used as a tool to measure a person's knowledge, skills or abilities. It is important to remember that every individual is unique and has different needs. Some people may have learning tendencies or ways of thinking that are incompatible with certain test formats, which may result in results that do not fully reflect their abilities. In addition, there are also other factors such as anxiety, stress or mental health disorders that can affect a person's performance in tests.

[3] P.J. Surkan et al., "A Qualitative Evaluation of the Use of Problem Management Plus (PM+) among Arabic-Speaking Migrants with Psychological Distress in France–The APEX Study," *European Journal of Psychotraumatology* 15, no. 1 (2024), https://doi.org/10.1080/20008066.2024.2325243.

[4] Suddin Bani, "OBJEK EVALUASI PENDIDIKAN," *Lentera Pendidikan : Jurnal Ilmu Tarbiyah dan Keguruan* 15, no. 2 (December 20, 2012): 231–39, https://doi.org/10.24252/lp.2012v15n2a8.

[5] Barney Glaser and Anselm Strauss, *Discovery of Grounded Theory: Strategies for Qualitative Research* (Routledge, 2017),

Therefore, it is important to consider a variety of evaluation methods, including performance assessments, portfolios, or direct observation, to ensure that all individuals have a fair chance to demonstrate their all-round understanding and abilities.[6]

Previous research, conducted by Ana Ratwa Wulan[7] explored the essence of the concepts of assessment, testing and measurement. The desired study results in evaluation for continuous improvement of lessons for students and learning outcomes are stated based on real data obtained by conducting in-depth and thorough assessments so that the data can be seen accurately and precisely. Therefore, there are several terms related to tests, namely testing and measurement, which are often used by teachers and teachers.[8]

In addition, research by Indah Rahmi et.al.[9] reviews the analysis of the quality of Arabic tests based on Higher Order Thinking Skill (HOTS) in the realm of evaluation which is very minimal in MTs Al-Musyawarah Lembang. In this case, an analysis was carried out to determine the quality of Arabic End-of-Semester

---

[6] Radha Mohan, *Measurement, Evaluation and Assessment in Education* (PHI Learning Pvt. Ltd., 2023),

[7] Ana Ratna Wulan, "Pengertian Dan Esensi Konsep Evaluasi, Asesmen, Tes, Dan Pengukuran," *Jurnal, FPMIPA Universitas Pendidikan Indonesia*, 2007, https://www.academia.edu/download/34534033/pengertian_asesmen.pdf.

[8] Da'ad Abdel-Hay et al., "The Arabic EAT-10 and FEES in Dysphagia Screening among Cancer Patients: A Comparative Prospective Study," *Scientific Reports* 14, no. 1 (April 22, 2024): 9258, https://doi.org/10.1038/s41598-024-58572-z.

[9] Indah Rahmi Nur Fauziah, Syihabudin Syihabudin, and Asep Sopian, "Analisis Kualitas Tes Bahasa Arab Berbasis Higher Order Thinking Skill (Hots)," *لساننا (Lisanuna): Jurnal Ilmu Bahasa Arab Dan Pembelajarannya* 10, no. 1 (2020): 45–54.

Test (UAS) items by conducting descriptive research with a total sample of 30 people. The conclusion of this research shows that test validity and reliability have a significant value, 25 questions do not meet the criteria in terms of multiple choice writing rules. As for the level of difficulty obtained, there is no suitability, the differential power is medium, and the effectiveness of the triggers is sufficient.

Furthermore, research conducted by Dina Indriana[10] describes learning evaluation and authentic assessment in Arabic language learning. Evaluation is carried out by the CIPP model to determine the extent of learning outcomes by using an approach to 2 important concepts in the world of education that refer to the development of deep understanding and critical thinking skills of students. This strategy is part of learning and assessment which becomes a benchmark in implementing learning strategies that have been carried out by teachers and students[11].

These three previous studies provide the latest scientific information related to the analysis of test evaluation in the proposed scientific and authentic approach to see the extent to which the test evaluation results are carried out accurately and precisely.[12] In this

[10] Dina Indriana, "Evaluasi Pembelajaran Dan Penilaian Autentik Dalam Pembelajaran Bahasa Arab," *Al-Ittijah: Jurnal Keilmuan Dan Kependidikan Bahasa Arab* 10, no. 2 (2018): 34–52.

[11] A.A. Alghamdi et al., "The Translation and Validation of the Surgical Anxiety Questionnaire into the Modern Standard Arabic Language: Results from Classical Test Theory and Item Response Theory Analyses," *BMC Psychiatry* 24, no. 1 (2024), https://doi.org/10.1186/s12888-024-06142-y.

[12] Amy Fitriani Siregar et al., "Test Analysis of Durūs Al-Lughah Al-'Arabiyyah Volume 1 by Imam Zarkasyi and Imam Syubani," *Lisanudhad: Jurnal Bahasa,*

case the researcher conducted a multiple choice modeled test in Arabic language learning in college. By conducting feasibility and reliability evaluation tests on multiple choice, through this approach, it provides an overview and offers that can be the right solution to deepen students' understanding of Arabic language learning concepts. Through this research, it is expected to find feasibility and reliability results that are relevant and accurate to improve the quality of Arabic language learning.[13]

Method

This research uses a quantitative approach with a test validation study design to evaluate the feasibility and reliability of Arabic multiple choice tests at UIN Sunan Ampel Surabaya. Data were collected through administering the multiple choice test to student participants at various levels of higher education. In this process, validity and reliability analyses were conducted to assess the extent to which the test appropriately measures Arabic language proficiency and the consistency of test results over time.[14] In line with this research, "Test validity and reliability are crucial in

Pembelajaran, Dan Sastra Arab 11, no. 01 (June 25, 2024): 153–75, https://doi.org/10.21111/lisanudhad.v11i01.11427.

[13] Dima Bitar and Marie Oscarsson, "Arabic-Speaking Women's Experiences of Communication at Antenatal Care in Sweden Using a Tablet Application—Part of Development and Feasibility Study," *Midwifery* 84 (2020): 102660.

[14] R. Magdy et al., "Validity and Reliability of Arabic Version of Pediatric Migraine Disability Assessment Scale (Child Self-Report versus Parent Proxy-Report): A Multi-Center Study," *Journal of Headache and Pain* 25, no. 1 (2024), https://doi.org/10.1186/s10194-024-01713-6.

ensuring the integrity and effectiveness of test use in the context of higher education.[15]

Furthermore, the use of validity tests such as the correlation between test results and student academic performance can provide an understanding of the extent to which the test can be trusted as an accurate assessment tool. By comparing test scores with students' academic achievements in other Arabic courses, the research can assess the extent to which the test really measures students' Arabic language skills well.[16]

To evaluate test reliability, the research will use internal reliability testing methods such as Cronbach's alpha to measure the internal consistency between test items. This step will help to identify whether the items in the test are related and measure the same concept consistently.[17] Furthermore, to measure the reliability of the test, the research will conduct a retest on the same sample of students to see how consistent the test results are over time. By combining these two methods, the research can provide a holistic

[15] Muncar Winarti, Abdurrachman Faridi, and Fahrur Rozi, "Evaluating the Validity, Reliability and Authenticity of English Achievement Test for the Twelfth Grade Students of SMAN 4 Tebo, Jambi," *English Education Journal* 11, no. 1 (March 15, 2021): 130–38, https://doi.org/10.15294/eej.v11i1.44176.

[16] Itsna Oktaviyanti and N. K. R. Awal, "Korelasi Antara Hasil Tes Lisan Dengan Hasil Tes Tertulis Pada Mahasiswa PGSD UNRAM," *Jurnal Ilmu Pendidikan* 2, no. 1 (2019): 9–19.

[17] Sionara Tamanini de Almeida, Thais de Lima Resende, and Claus Dieter Stobäus, "Validity, Reliability and Convergent Analysis of Brazilian Version of Selection, Optimization and Compensation Questionnaire (QSOC)," *Creative Education* 7, no. 15 (September 6, 2016): 2074–87, https://doi.org/10.4236/ce.2016.715207.

understanding of the reliability and consistency of Arabic multiple choice tests used in the college environment.

Result and Discussion

Test Content analysis

From an analysis of 25 Arabic multiple choice questions used at UIN Sunan Ampel Surabaya, it was found that 60% of the questions focused on basic grammar, 25% on vocabulary, and only 15% on text comprehension. Grammar questions generally assess students' abilities in the context of simple sentences, while vocabulary questions more often assess the recognition and use of basic words. The text comprehension aspect, which is an important part of language competence, seems to be underrepresented in the test, with only 15% of questions testing the ability to analyze and understand more complex texts.[18]

The results of this analysis suggest that the Arabic multiple choice test at UIN Sunan Ampel Surabaya needs to be further evaluated to ensure it is more comprehensive in its coverage of all aspects of the Arabic teaching curriculum. Currently, the test tends to emphasize more on basic grammar and vocabulary, while aspects of text comprehension and language use in more complex contexts are underrepresented.[19] This evaluation and adjustment is important

[18] Brightlin Nithis Dhas et al., "Psychometric Properties of the Arabic Occupational Balance Questionnaire (OBQ11-A)," *Annals of Medicine* 56, no. 1 (December 31, 2024): 2346945, https://doi.org/10.1080/07853890.2024.2346945.

[19] Elana Shohamy, Iair G. Or, and Stephen May, *Language Testing and Assessment* (Springer Cham, 2017),

so that each item in the test reflects the competencies expected from the students, such as the ability to analyze texts and use language in diverse realistic situations. By making these adjustments, multiple choice tests will be more representative of the curriculum and able to measure students' competencies more thoroughly and accurately.[20]

It is important to integrate more varied items in testing grammar usage in complex sentence contexts, so as to reflect students' abilities more accurately and in depth. According to Nunnally (1978), tests that have good internal consistency and cover various aspects of skills can improve their reliability and validity.[21] Thus, this multiple choice test must be updated regularly to ensure its relevance to curriculum development and student learning needs at UIN Sunan Ampel Surabaya.

Test Validity

In assessing a multiple choice test or seeing the quality of the test whether it is good or not, there are things that must be fulfilled so that the test is of high quality. Each test will be checked for quality in four categories as follows: validity, reliability, power differential, and difficulty level. Validity ensures that the test

[20] H. Douglas Brown and Priyanvada Abeywickrama, "Language Assessment," *Principles and Classroom Practices. White Plains, NY: Pearson Education*, 2004, 20.

[21] J. C. Nunnally, "Psychometric Theory 2nd Edition (New York: McGraw)," 1978.

measures what it is supposed to measure[22] reliability ensures the consistency of the test results, differentiability determines the test's ability to distinguish between students who have different knowledge, and difficulty level ensures the questions have a variety of difficulties that match the students' ability levels.[23] By fulfilling these four categories, multiple choice tests can be said to be of high quality and able to provide an accurate assessment of students' Arabic language skills.[24]

Validity is a measure that indicates the extent to which a test measures what it is supposed to measure. In the context of educational evaluation, validity ensures that the test is able to measure the ability or knowledge expected of learners in accordance with the learning objectives. According to Messick (1989), validity is the most critical aspect of a test because without validity, test results cannot be used to make accurate decisions about test takers' abilities. Test validity includes various types, including content

[22] Cátia Quintão, Pedro Andrade, and Fernando Almeida, "How to Improve the Validity and Reliability of a Case Study Approach?," *Journal of Interdisciplinary Studies in Education* 9, no. 2 (2020): 264–75.

[23] S. Alsaleh et al., "Reliability and Validity of the Arabic Version of the Brief Version of the Questionnaire of Olfactory Disorders," *Laryngoscope Investigative Otolaryngology* 9, no. 3 (2024), https://doi.org/10.1002/lio2.1252.

[24] Mustapha Qureshi et al., "Scale for Measuring Arabic Speaking Skills in Early Children's Education.," *JILTECH: Journal International of Lingua & Technology* 1, no. 2 (2022),

validity, construct validity, and criterion validity, all of which contribute to the accuracy and relevance of test results.[25]

Display validity is concerned with the general perception of the measuring instrument.[26] In the context of validity discussed earlier, display validity is not properly categorized as a type of validity because it is not directly related to the ability of the measuring instrument to measure what it is supposed to measure. Display validity may be less scientific and based more on habit or perception, such as the way choices are arranged in multiple-choice questions, than on empirical or theoretical evidence.[27] The following table shows the assessment of display validity by teachers and students:

Table 1. Display Validity Assessment by Teachers and Students

No	Aspect of question	Teachers (0%) who agree	Students (0%) who agree
1	Relevant to curriculum materials	85%	90%
2	Clear and precise grammar	90%	88%

[25] S. Messick, "Validity. Em r. Linn (Org.), Educational Measurement.(13-103)," *New York, NY: American Council on Education and Macmillan Publishing Company*, 1989.

[26] Intan Deviana Ilyas, "PENGUJIAN USABILITY WEBSITE SMKN 1 PONOROGO MENGGUNAKAN SYSTEM USABILITY SCALE" (PhD Thesis, Universitas Muhammadiyah Ponorogo, 2020), http://eprints.umpo.ac.id/6236/.

[27] Rasa Jankauskiene, Danielius Urmanavicius, and Migle Baceviciene, "Associations between Perceived Teacher Autonomy Support, Self-Determined Motivation, Physical Activity Habits and Non-Participation in Physical Education in a Sample of Lithuanian Adolescents," *Behavioral Sciences* 12, no. 9 (2022): 314.

3	Interesting multiple choice format	75%	82%
4	Questions cover a wide range of competencies	60%	55%

Table 1 shows the results of the display validity assessment from teachers and students. Most of the teachers (85%) and students (80%) agreed that the test questions were relevant to the curriculum materials. This high percentage indicates that the test has good display validity according to their perception. That is, in general, teachers and students felt that the test looked appropriate to what was taught and learned in class, although this judgment was based more on subjective perception than empirical analysis.

However, there were differences in perceptions regarding the extent to which the questions covered the range of competencies expected. Only 60% of teachers and 55% of students agreed that the questions covered the expected range of competencies. This suggests that although the tests appear relevant to the curriculum, there is a concern that the items may not be varied enough to measure all the desired aspects of Arabic language proficiency. Therefore, despite the high face validity, there is still a need to improve and refine the questions to be more representative of all the competencies taught in the curriculum.

The empirical validity of the Arabic multiple choice test showed that only 60% of the questions were in line with the curriculum, and of the 25 questions analyzed, only 14 questions were empirically valid. The reliability value of the test, measured by Cronbach's Alpha, was 0.88, indicating fairly good internal

consistency. In addition, the correlation between test results and students' academic performance was r = 0.44, indicating a significant but not very strong positive correlation. These results indicate that although the test has some aspects of validity, further improvements are still needed to enhance its suitability to the curriculum and accuracy in measuring students' overall Arabic language competence. The following table shows the Empirical Validity of the Arabic Multiple Choice Test:

Table 2. Empirical validity of Arabic multiple choice test

No.	Aspects of Empirical Validity	Empirical Analysis Results
1	Content validity	60% of questions match the curriculum
2	Construct validity	14 out of 25 questions are valid
3	Reliability (Cronbach's Alpha)	0.88
4	Correlation with academic performance	r = 0.44

Table 2 shows the empirical validity of the Arabic multiple choice test. This validity shows that only 60% of the items were curriculum compliant, and of the 25 items analysed, only 14 items were empirically valid. The reliability value of the test, measured by Cronbach's Alpha, is 0.88, which indicates excellent internal consistency. In addition, the correlation between the test results and students' academic performance was r = 0.44, indicating a significant but not very strong positive correlation.

These results indicate that although the test has some good aspects of validity, further improvements are needed to enhance its fit with the curriculum. This improvement is important so that the test can more accurately measure students' overall Arabic language competence [28]. By fixing the invalid questions and ensuring that all aspects of the curriculum are covered thoroughly, this multiple choice test will be able to provide a more precise and comprehensive assessment of students' Arabic language skills.

However, out of the 25 multiple choice questions analyzed, only 14 questions proved to be valid in effectively measuring Arabic language competence. This suggests that there is room for improvement in item design to be more representative of the curriculum and expected learning objectives. By improving the validity of the less effective questions, the test can provide a more accurate picture of students' Arabic language ability and be used as a more comprehensive and appropriate evaluation tool.[29] According to Brown & Abeywickrama (2004), a valid test is one that is able to reflect the ability it is intended to measure, and the results of this test show that there are still steps that need to be taken to achieve an optimal level of validity.[30]

However, it should be noted that some item designs still need to be improved to increase the overall representation of the

[28] Muhammad Baihaqi, "Evaluasi Pembelajaran," *Surabaya: LAPIS PGMI*, 2008.
[29] Farida Far Ida and Anna Musyarofah, "Validitas Dan Reliabilitas Dalam Analisis Butir Soal," *Al-Muarrib : Jurnal Pendidikan Bahasa Arab* 1, no. 1 (December 6, 2021): 34–44, https://doi.org/10.32923/al-muarrib.v1i1.2100.
[30] Brown and Abeywickrama, "Language Assessment."

curriculum and ensure that the test remains relevant to the expected learning objectives. Although the multiple choice test has shown good validity, only 14 out of 25 items proved to be effective in measuring students' overall Arabic language competence. This indicates that some questions focus too much on certain aspects such as basic grammar, while other aspects such as comprehension of complex texts and language use in realistic contexts are underrepresented. By improving the question design to be more balanced and cover various aspects of the curriculum, the test will be able to provide a more accurate and comprehensive assessment of students' Arabic language skills, in accordance with the evaluation principles advocated by Erna Wurjanti.[31]

Outward validity is a type of validity that refers to the extent to which a research instrument or measurement appears appropriate or relevant for the purpose or concept being researched.[32] Focuses on the subjective judgement of the person using the instrument whether the instrument seems to measure what it is supposed to measure. For example, if a researcher designs a questionnaire to measure customer satisfaction, the outward validity of the questionnaire will be evaluated by how well it appears to measure customer satisfaction in the view of the person completing the questionnaire.

---

[31] Erna Wurjanti, *Study Group Solusi Meningkatkan Motivasi Dan Hasil Belajar* (Penerbit P4I, 2022),

[32] Andri Wicaksono, *Metodologi Penelitian Pendidikan: Pengantar Ringkas* (Garudhawaca, 2022),

In this regard, outward validity can be a first step in evaluating an instrument, but it is not sufficient to guarantee its overall validity. This is because subjective judgements can be influenced by many factors and may not reflect the extent to which the instrument actually measures what is intended. Therefore, it is important to also consider other types of validity, such as construct validity or criterion validity, to assess whether the instrument actually measures what is scientifically intended. The following table shows the external validity of the Arabic multiple choice test:

Table 3. External Validity of Arabic Multiple Choice Test

No.	Number of Test	The Available Questions	Should be	Reason
1	4	... مدير الجامعة ١. هذه ٢. تلك ٣. هذا ٤. نحن	هذه هو هذا نحن	None of the answer choices are correct, because "مدير الجامعة" means university director. The correct one uses the dhomir "هو", so the "تلك" is replaced by "هو".
2	12	ذهبت ... إلى الجامعة صباح أ. الطلاب	أ. الطلاب ب. خديجة ج. المدرّسون د. العمّال	None of the answer choices are correct, the answer should be خديجة without وفائزة because

ذهبت is a madhi fi'il referring to هي		ب. خديجة وفائزة ج. المدرّسون د. العمّال		

From table 3, several errors were found in the answer choices that require correction. In question 4, the correct answer choice should be "**هو**" because "**مدير الجامعة**" means university director, which refers to the dhomir "**هو**". The option "**تلك**" was replaced with "**هو**" to correct the error. In question 12, the correct answer should be "**خديجة**" without "**وفائزة**" because "**ذهبت**" is a madhi fi'il that refers to "**هي**". In the answer options given, none of them are correct, so the option "**خديجة**" should be adjusted by omitting "**وفائزة**" to ensure conformity with Arabic grammar rules**.**

However, it should be noted that some item designs still need to be revised to improve the overall face validity and ensure that the test remains relevant to the expected learning objectives. Revisions are needed to ensure that each item reflects the expected competencies and covers all aspects of the material taught, so that the test can provide an accurate and comprehensive assessment of students' Arabic language skills**.**

Construct validity is a measure that indicates the extent to which a test measures the theoretical concept or construct that it is

supposed to measure [33]. This validity ensures that the test actually measures the intended characteristics or abilities, not other irrelevant things. In the context of Arabic language tests, construct validity will ensure that the test questions actually measure Arabic language abilities, such as grammar comprehension, vocabulary, and reading skills, in accordance with the theory or model of language ability underlying the curriculum. The following table shows the construct validity of multiple choice tests:

Table 4. Arabic Multiple Choice Test Construct Validity

No.	Number of Test	The Available Questions	Should be	Reason
1	2	المسجد . . . . أ. ماهر ب. صغرية ج. سريعة د. كبير	المسجد . . . أ. ثقيل ب. صغرية ج. سريعة د. كبير	In the question, four dots should have been used. The answer choices are not homogeneous and do not work as an exemption. It should be replaced with an adjective.
2	3	اهتم العلماء بالحديث النبوي اهتماما. ...	اهتم العلماء بالحديث	In the question, there are four points and three should have been used.

[33] Wenda Asmita and Wahidah Fitriani, "KONSEP DASAR PENGUKURAN," *Jurnal Mahasiswa BK An-Nur: Berbeda, Bermakna, Mulia* 8, no. 3 (2022): 217–26.

	النبوي اهتماما...			
In the question, there are four points and three should have been used.	نشأ هامكا في عائلة مسلمة متدينة في سوماطرا . . .	نشأ هامكا في عائلة مسلمة متدينة في سوماطرا . . .	11	3
In the question, there are four points and three should have been used.	إندونيسيا مناظرها . . .	إندونيسيا مناظرها	15	4
The question does not end with a full stop. Each sentence should end with a full stop.	. . . فحصتا المريض .	. . . فحصتا المريض	18	5
The question uses an exclamation mark, but it should use a question mark because the sentence is a question.	ما هو القرآن الكريم؟	ما هو القرآن الكريم!	20	6

Table 4 shows the construct validity analysis of the Arabic multiple choice test with a focus on errors in question writing and design. In question 2, it was found that there were four dots in the sentence "المسجد . . . ." three dots should have been used. In addition,

the answer choices were not homogeneous and effective as exemptions, so they needed to be replaced with more appropriate adjectives. Problem numbers 3 and 11 also suffer from similar problems, where the use of four dots should have been changed to three dots to maintain consistency and clarity. This small error can affect the construct validity of the test as it interferes with the clarity and interpretation of the question by test takers.

Furthermore, questions 18 and 20 do not end with a full stop, whereas every sentence should end with a full stop to maintain grammatical uniformity and clarity. Questions in the writing section, such as "ما هو القرآن الكريم" should use a question mark instead of an exclamation mark, because it is a question sentence. In addition, in the question "وسائل النقل العامة أوالخاصة، أيهما تركب عندما ذهبت إلى الجامعة؟ ولماذا؟" there is a layout error where after the word "أو" there should be a space so that it does not merge with the next word. Correcting these errors is essential to ensure that the test measures the intended theoretical construct accurately and appropriately, so that the construct validity of the test can be maintained and improved.

Test Reliability

Test reliability refers to the extent to which a test is consistent and reliable in measuring what it purports to measure without

significant measurement error.[34] It reflects the degree of stability or consistency of the results obtained from the test over time, as well as how well the test can produce consistent scores for the same subject if tested multiple times. In other words, test reliability measures how accurate and consistent the test is in measuring what it is supposed to measure without significant variability due to external factors or measurement error. Evaluation of test reliability is important to ensure that test results are reliable and can be trusted in making decisions or inferences about the individuals or phenomena measured by the test.

The Cronbach's alpha value of 0.88 indicates that the multiple choice test has excellent internal consistency, which means that the items in the test have a high level of reliability and consistently measure the same construct. This indicates that the test can be trusted to provide stable and accurate results each time it is used, making it an effective evaluation tool for measuring students' Arabic language proficiency.

Although the Cronbach's Alpha value obtained is greater than 0.70, indicating an adequate level of reliability, the conclusion that the data used is reliable should be made with caution. Although Cronbach's Alpha is a commonly used method for measuring the internal reliability of an instrument, the validity of the interpretation of Alpha values must take into account the specific context of the instrument and the population under study. A holistic assessment of

[34] Anne Anastasi and Susana Urbina, "Psychological Testing 7th Ed. Prentics-Hall International" (Inc, 1997).

other aspects of reliability, such as the stability and consistency of the instrument, as well as consideration of whether the instrument actually measures the intended construct, is important to ensure accurate conclusions about the reliability of the data.

Conclusion

Based on the results of the study, it was found that multiple choice tests in Arabic in higher education have a fairly high content validity, with 60% of the questions in accordance with the existing curriculum. However, its construct validity still needs to be improved as only 14 out of 25 questions are considered valid. However, the reliability of the test measured using Cronbach's Alpha showed a high level of reliability at 0.88. There was also a significant positive correlation between test scores and students' academic performance, although the correlation was not very strong ($r = 0.44$). Therefore, despite the shortcomings in construct validity, the test is still considered reliable as an evaluation tool, especially in the context of measuring students' academic performance. This conclusion provides an overall picture of the effectiveness of the test in measuring students' understanding and mastery of the Arabic language, as well as how reliable the test results are in making academic evaluation decisions.

References

Abdel-Hay, Da’ad, Osama Abdelhay, Hamza A. Ghatasheh, Sameer Al-Jarrah, Suhaib Eid, Mutaz A. Al Tamimi, and Ibrahim Al-Mayata. “The Arabic EAT-10 and FEES in Dysphagia

Screening among Cancer Patients: A Comparative Prospective Study." *Scientific Reports* 14, no. 1 (April 22, 2024): 9258. https://doi.org/10.1038/s41598-024-58572-z.

Alghamdi, A.A., K. Alghuthayr, S.S.S.M.M. Alqahtani, Z.A. Alshahrani, A.M. Asiri, H. Ghazzawi, M. Helmy, K. Trabelsi, M. Husni, and H. Jahrami. "The Translation and Validation of the Surgical Anxiety Questionnaire into the Modern Standard Arabic Language: Results from Classical Test Theory and Item Response Theory Analyses." *BMC Psychiatry* 24, no. 1 (2024). https://doi.org/10.1186/s12888-024-06142-y.

Almeida, Sionara Tamanini de, Thais de Lima Resende, and Claus Dieter Stobäus. "Validity, Reliability and Convergent Analysis of Brazilian Version of Selection, Optimization and Compensation Questionnaire (QSOC)." *Creative Education* 7, no. 15 (September 6, 2016): 2074–87. https://doi.org/10.4236/ce.2016.715207.

Alsaleh, S., R. Alfallaj, H. Almousa, N. Alsubaie, Y. Akkielah, T.A. Mesallam, and I. Sumaily. "Reliability and Validity of the Arabic Version of the Brief Version of the Questionnaire of Olfactory Disorders." *Laryngoscope Investigative Otolaryngology* 9, no. 3 (2024). https://doi.org/10.1002/lio2.1252.

Anastasi, Anne, and Susana Urbina. "Psychological Testing 7th Ed. Prentics-Hall International." Inc, 1997.

Asmita, Wenda, and Wahidah Fitriani. "KONSEP DASAR PENGUKURAN." *Jurnal Mahasiswa BK An-Nur: Berbeda, Bermakna, Mulia* 8, no. 3 (2022): 217–26.

Baihaqi, Muhammad. "Evaluasi Pembelajaran." *Surabaya: LAPIS PGMI*, 2008.

Bani, Suddin. "OBJEK EVALUASI PENDIDIKAN." *Lentera Pendidikan : Jurnal Ilmu Tarbiyah dan Keguruan* 15, no. 2 (December 20, 2012): 231–39. https://doi.org/10.24252/lp.2012v15n2a8.

Bitar, Dima, and Marie Oscarsson. "Arabic-Speaking Women's Experiences of Communication at Antenatal Care in Sweden Using a Tablet Application—Part of Development and Feasibility Study." *Midwifery* 84 (2020): 102660.

Brown, H. Douglas, and Priyanvada Abeywickrama. "Language Assessment." *Principles and Classroom Practices. White Plains, NY: Pearson Education*, 2004, 20.

Deviana Ilyas, Intan. "PENGUJIAN USABILITY WEBSITE SMKN 1 PONOROGO MENGGUNAKAN SYSTEM USABILITY SCALE." PhD Thesis, Universitas Muhammadiyah Ponorogo, 2020. http://eprints.umpo.ac.id/6236/.

Dhas, Brightlin Nithis, Samah Ahmad Abd Alfattah Abd Alhadi, Ghaith Mohammad Rizk Dhadl Al That, and Sultan Salim Hammam Al Abdulla. "Psychometric Properties of the Arabic Occupational Balance Questionnaire (OBQ11-A)." *Annals of Medicine* 56, no. 1 (December 31, 2024): 2346945. https://doi.org/10.1080/07853890.2024.2346945.

Fathi, Jalil, Lawrence Jun Zhang, and Mohammad Hossein Arefian. "Testing a Model of EFL Teachers' Work Engagement: The

Roles of Teachers' Professional Identity, L2 Grit, and Foreign Language Teaching Enjoyment." *International Review of Applied Linguistics in Language Teaching* 0, no. 0 (July 21, 2023). https://doi.org/10.1515/iral-2023-0024.

Fauziah, Indah Rahmi Nur, Syihabudin Syihabudin, and Asep Sopian. "Analisis Kualitas Tes Bahasa Arab Berbasis Higher Order Thinking Skill (Hots)." *لسانـنا (Lisanuna): Jurnal Ilmu Bahasa Arab Dan Pembelajarannya* 10, no. 1 (2020): 45–54.

Glaser, Barney, and Anselm Strauss. *Discovery of Grounded Theory: Strategies for Qualitative Research*. Routledge, 2017. https://www.taylorfrancis.com/books/mono/10.4324/9780203793206/discovery-grounded-theory-barney-glaser-anselm-strauss.

Ida, Farida Far, and Anna Musyarofah. "Validitas Dan Reliabilitas Dalam Analisis Butir Soal." *Al-Muarrib : Jurnal Pendidikan Bahasa Arab* 1, no. 1 (December 6, 2021): 34–44. https://doi.org/10.32923/al-muarrib.v1i1.2100.

Indriana, Dina. "Evaluasi Pembelajaran Dan Penilaian Autentik Dalam Pembelajaran Bahasa Arab." *Al-Ittijah: Jurnal Keilmuan Dan Kependidikan Bahasa Arab* 10, no. 2 (2018): 34–52.

Jankauskiene, Rasa, Danielius Urmanavicius, and Migle Baceviciene. "Associations between Perceived Teacher Autonomy Support, Self-Determined Motivation, Physical Activity Habits and Non-Participation in Physical Education in a Sample of Lithuanian Adolescents." *Behavioral Sciences* 12, no. 9 (2022): 314.

Magdy, R., A. Hassan, Z. Mohammed, M.A. Abdeltwab, N.F.A. Ghaffar, and M. Hussein. "Validity and Reliability of Arabic Version of Pediatric Migraine Disability Assessment Scale (Child Self-Report versus Parent Proxy-Report): A Multi-Center Study." *Journal of Headache and Pain* 25, no. 1 (2024). https://doi.org/10.1186/s10194-024-01713-6.

Messick, S. "Validity. Em r. Linn (Org.), Educational Measurement.(13-103)." *New York, NY: American Council on Education and Macmillan Publishing Company*, 1989.

Mohan, Radha. *Measurement, Evaluation and Assessment in Education*. PHI Learning Pvt. Ltd., 2023.

Nunnally, J. C. "Psychometric Theory 2nd Edition (New York: McGraw)," 1978.

Oktaviyanti, Itsna, and N. K. R. Awal. "Korelasi Antara Hasil Tes Lisan Dengan Hasil Tes Tertulis Pada Mahasiswa PGSD UNRAM." *Jurnal Ilmu Pendidikan* 2, no. 1 (2019): 9–19.

Quintão, Cátia, Pedro Andrade, and Fernando Almeida. "How to Improve the Validity and Reliability of a Case Study Approach?" *Journal of Interdisciplinary Studies in Education* 9, no. 2 (2020): 264–75.

Qureshi, Mustapha, Dinnah Mahdiyyah, Yassine Mohamed, and Mounika Ardchir. "Scale for Measuring Arabic Speaking Skills in Early Children's Education." *JILTECH: Journal International of Lingua & Technology* 1, no. 2 (2022).

Safdari, Maryam, and Jalil Fathi. "Investigating the Role of Dynamic Assessment on Speaking Accuracy and Fluency of

Pre-Intermediate EFL Learners." Edited by Richard Kruk. *Cogent Education* 7, no. 1 (January 1, 2020): 1818924. https://doi.org/10.1080/2331186X.2020.1818924.

Shohamy, Elana, Iair G. Or, and Stephen May. *Language Testing and Assessment*. Springer Cham, 2017.

Siregar, Amy Fitriani, Siti Nurhasana Mokodompit, Muhajir Muhajir, and Nila Alfiroh. "Test Analysis of Durūs Al-Lughah Al-'Arabiyyah Volume 1 by Imam Zarkasyi and Imam Syubani." *Lisanudhad: Jurnal Bahasa, Pembelajaran, Dan Sastra Arab* 11, no. 01 (June 25, 2024): 153–75. https://doi.org/10.21111/lisanudhad.v11i01.11427.

Surkan, P.J., D. Rayes, L. Bertuzzi, N. Figueiredo, M. Melchior, and A. Tortelli. "A Qualitative Evaluation of the Use of Problem Management Plus (PM+) among Arabic-Speaking Migrants with Psychological Distress in France–The APEX Study." *European Journal of Psychotraumatology* 15, no. 1 (2024). https://doi.org/10.1080/20008066.2024.2325243.

Wicaksono, Andri. *Metodologi Penelitian Pendidikan: Pengantar Ringkas*. Garudhawaca, 2022.

Winarti, Muncar, Abdurrachman Faridi, and Fahrur Rozi. "Evaluating the Validity, Reliability and Authenticity of English Achievement Test for the Twelfth Grade Students of SMAN 4 Tebo, Jambi." *English Education Journal* 11, no. 1 (March 15, 2021): 130–38. https://doi.org/10.15294/eej.v11i1.44176.

Wulan, Ana Ratna. "Pengertian Dan Esensi Konsep Evaluasi, Asesmen, Tes, Dan Pengukuran." *Jurnal, FPMIPA*

Universitas Pendidikan Indonesia, 2007. https://www.academia.edu/download/34534033/pengertian_asesmen.pdf.

Wurjanti, Erna. *Study Group Solusi Meningkatkan Motivasi Dan Hasil Belajar*. Penerbit P4I, 2022.