Edulab :
Majalah Ilmiah Laboratorium Pendidikan
https://ejournal.uin-suka.ac.id/tarbiyah/index.php/edulab/index

EISSN: 2527-7200
Vol. 9, No. 2, December 2024
DOI: https://doi.org/10.14421/edulab.2024.92.01

Analysis of Item Difficulty Levels and Differentiating Power of TOAFL Questions at KH. Abdul Wahab Hasbullah University Jombang

Bustanil Ilmi Agustin ✉
UIN Sunan Ampel Surabaya, Indonesia
Rikha Ikke Nuriani
Universitas Negeri Malang, Indonesia
Nur Qurrotaa'yun
UIN Sunan Ampel Surabaya, Indonesia
Baihaqi
UIN Sunan Ampel Surabaya, Indonesia

ABSTRACT

Purpose – This study aims to evaluate the quality of the Test of Arabic Foreign Language (TOAFL) items administered at KH. Abdul Wahab Hasbullah University (UNWAHA), Jombang. Since the onset of the COVID-19 pandemic, the standard for TOAFL questions has been lowered, making them easier. Therefore, it is necessary to analyze the item difficulty levels and differentiating power to assess the test's quality.

Design/methods/approach – This study employed a quantitative descriptive method. Data were obtained from the results of TOAFL Package A, administered to 20 UNWAHA students in 2024. Data collection was conducted through documentation and analyzed using ANATES software.

Findings – The findings revealed that among 140 questions, the level of difficulty of UNWAHA TOAFL items was considered good because it had a balanced proportion, namely 13 items (9.3%) easy category, 83 items (59.3%) moderate category and 44 items (31.5%) difficult category. The differentiating power of the test questions is also said to be good (56.4%) because more than 50% of the questions are able to distinguish between upper and lower group students. Only a small proportion had less (12.8%) and negatif (14.3%) differentiating power. Based on these results, the questions can be reused in future test.

Research implications/limitations – This research can provide important input for the TOAFL question-compiling team to improve the quality of questions for the better. However, this reaearch is limited to one question package only.

Originality/value – This research discusses all language skills on the TOAFL test, which is tested on students from various majors and then analyzed using software.

OPEN ACCESS

ARTICLE HISTORY
Received: 30-09-2024
Revised: 23-11-2024
Accepted: 15-12-2024

KEYWORDS
Differentiating Power, Evaluation, Level of Difficulty, TOAFL

CONTACT: ✉02040923004@uinsa.ac.id

Introduction

Arabic is considered a foreign language in countries where it is not formally used (Taufik et al., 2023). In Indonesia, Arabic is taught in various educational institutions, both formal and non-formal, from elementary to higher education levels (Susiawati et al., 2022). Learning Arabic poses greater challenges compared to learning regional languages or Indonesian (Hamdun & Islam, 2023) Various challenges emerge during the learning process, including those related to students, teachers, curricula, and others. To identify and address these challenges, evaluation is essential for improving future learning processes (Bamualim, 2020)**.**

Evaluation is a crucial component in determining students' learning success. It is integral to the teaching and learning process (Muhimmatul Choiroh, 2021). Evaluation enables teachers to assess the achievement of educational objectives (Muhammad Lukman Arifianto et al., 2021). Tests are essential in evaluating learning outcomes, as they measure the extent to which material has been effectively conveyed to students. Test results serve as an indicator of the learning process's success and as a reflection for enhancing teaching quality.

The Test of Arabic Foreign Language (TOAFL) is an evaluative instrument used to assess students' proficiency in Arabic (Salam et al., 2023). This test designed to measures students' language skills based on clear and measurable standards, covering listening, reading, writing, and grammar analysis skills (Qodri, 2020). Listening skills (istima') involve matching heard words with text and summarizing audio content (Pranata, 2022). Reading skills (qira'ah) include identifying main ideas and summarizing texts (Ishak & Fitriyanti, 2020). Minewhile writing skills (kitabah) are tested by determining the wrong word in a sentence, sorting words or sentences, and determining words according to the grammatical rules. (Rathomi, 2020).

TOAFL is a standardized test administered to students at Islamic Universities (PTKIN and PTKIS) to evaluate their Arabic abillity (L. Qomariyah & Niswah, 2021). KH. Abdul Wahab Hasbullah University (UNWAHA), Jombang, one of PTIKIS requires all undergraduate students to take this test as a graduation requirement. This test compiled by Arabic lecture team then Developed by the Language and Computer Laboratory, the test consists of 140 multiple-choice questions to measure listening skills (*istima'*), writing skills (*kitabah*), grammar skills (*tarakib wa al ibarah al arabiyyah*), and reading skills (*qira'ah*).

However, the overall test questions are not yet known to have met the criteria of a quality test tool or not. So there is a possibility that students who do not reach the minimum TOAFL score are not due to lack of understanding of the material, but because the items of test do not meet the criteria for quality questions. The quality of a good items if it meets several criteria, including validity, consistency, difficulty level, and differentiating power (Ainin, 2016). Thus, analyzing the TOAFL items' quality is crucial.

Research on TOAFL items has been found, such as research conducted by Utami. Significantly, this study analyzed item validity, content validity, and difficulty level of TOAFL questions at IAIN Ponorogo based on Bloom's Taxonomy perspective (Utami, 2018). The research findings on TOAFL items at IAIN Ponorogo indicate that the item validity and content validity are high; however, the difficulty level of the items requires further attention. This is attributed to the use of operational terms that are not yet ideal, with 50% focused on C1 and C4, while C2, C3, C5, and C6 show a result of 0%.

Secondly, Harahap's research on the quality of TOAFL items at IAIN Curup examined aspects of difficulty level, differentiating power, validity, and reliability (Harahap, 2018). The study revealed that the difficulty level of TOAFL items at IAIN Curup falls into the good category, with 15 items categorized as easy, 71 items as mediium, and 54 items as difficult. However, the differentiating power was classified as low category. The validity analysis showed that more than 50% of the items were deemed valid, while the reliability was rated as high with a score of 0.87.

Thirdly, Nurhayati's study on TOAFL items at the Arabic Education Department of the Faculty of Tarbiyah and Teacher Training at UIN Alauddin Makassar (B., 2020), found that 53 items (66.25%) were valid, while 27 items (33.75%) were invalid. In terms of reliability, the items demonstrated high reliability with a score of 0.83. The difficulty levels showed that 26 items (32.5%) were difficult, 42 items (52.5%) were moderate, and 12 items (15%) were easy. The differentiating power revealed that 20 items (25%) had poor differentiating power, 37 items (46.25%) were categorized as fair, 15 items (18.75%) as good, 1 item (1.25%) as very good, and 7 items (8.75%) exhibited negative or poor differentiating power. Regarding distractor effectiveness, 34 items (42.5%) were categorized as very good, 17 items (21.25%) as good, 24 items (30%) as poor, and 5 items (6.25%) as very poor.

Fourthly, Qomariyah's study on the quality of TOAFL items at Hasyim Asy'ari University (UNHASY) Tebuireng Jombang focused on reading skills (qiro'ah) in terms of difficulty levels and differentiating power (L. Qomariyah, 2022). The study revealed that the difficulty levels were considered adequate, comprising 23 easy items, 22 moderate items, and 5 difficult items. Similarly, the differentiating power was assessed as sufficient, as more than half of the items successfully distinguished between high-performing and low-performing students. Specifically, 44 items were deemed suitable for reuse, while 6 items required revision.

Fifthly, Halomoan's study on the quality of TOAFL items at UIN Sultan Syarif Kasim Riau examined aspects of validity, reliability, difficulty levels, and differentiating power (Halomoan et al., 2022). The results showed that 35 items (25%) were valid, while 105 items (75%) were invalid. The reliability of the items reached a score of 0.856. In terms of difficulty levels, 40 items (28.57%) were categorized as easy, 56 items (40%) as moderate, and 44 items (31.43%) as difficult. The differentiating power analysis revealed that 36 items (25.71%) were categorized as very good, 21 items (15%) as good, and 83 items (59.29%) were categorized as poor (low).

Sixthly, Wulandari's study on the quality of TOAFL items at IAIN Metro Lampung, particularly focusing on reading skills (qiro'ah), assessed validity, reliability, difficulty levels, and distractor effectiveness (Wulandari, 2023). The findings indicated that 24 items were valid, while 21 items were invalid. The reliability score was 0.8, indicating a high correlation. In terms of difficulty levels, there were 3 items categorized as difficult, 39 items as moderate, and 8 items as easy. However, the differentiating power of 2 items was classified as very poor, necessitating replacement.

This study differs from previous research in several ways. Some prior studies focused solely on specific aspects (e.g., reading skills), employed manual analysis techniques, and used TOAFL test samples limited to students majoring in Arabic language studies. In contrast, this study examines all language skills assessed in the TOAFL and involves students from various academic disciplines, with the analysis conducted using specialized software. Furthermore, this research was motivated by the decline in question standards, which became easier due to the impact of the COVID-19 pandemic on TOAFL test items.

Previous studies demonstrate that researchers have shown a strong commitment to ensuring the quality of Arabic language tests, aiming to develop test items that genuinely represent a reliable assessment tool and effectively measure students' language proficiency. However, this study differs from prior research in several ways. Many previous studies focused solely on specific aspects (e.g., reading skills), relied on manual analysis techniques, and used TOAFL test samples exclusively from Arabic language students. While, this study examines all language skills tested in the TOAFL, involving students from various disciplines, and employs software for the analysis. Additionally, this research was motivated by a decline in test item standards, which became easier due to the impact of the COVID-19 pandemic on TOAFL questions.

Therefore, the author aims to analyze the items of the TOAFL (Test of Arabic Foreign Language) used to assess the Arabic language proficiency of students at UNWAHA Jombang, based on their difficulty levels and discriminatory power. Through this item analysis, it will be possible to identify questions that are classified as easy, moderate, and difficult for the students. Additionally, this analysis also serves to evaluate the discriminatory power of the items, which plays a role in distinguishing between students with high and low proficiency (Pradita & Megawanti, 2023). A good question is one that effectively differentiates students' abilities (Fahmi et al., 2022). The results of this analysis will provide valuable feedback for the language laboratory to improve and refine the TOAFL items, ensuring that the test is of higher quality and standardized.

Methods

This research is a quantitative descriptive study. The data source used is the TOAFL Package A results from 20 students at UNWAHA in 2024. These students come from various faculties, including the Faculty of Education, Faculty of Economics, and Faculty of Islamic Studies. The sampling technique used is purposive sampling. This method was selected based on specific criteria with a relatively small number of participants to enhance efficiency. The study population includes all 140 items of the TOAFL, consisting of 50 listening questions (istima'), 40 language structure questions, writing (kitabah), and expressions in Arabic (tarakib wa al-ibarah al-arabiyyah), as well as 50 reading questions (qira'ah). The data source for this research consists of the TOAFL scores of the UNWAHA students and the frequency of correct answers for each item, which were collected through documentation. The collected data were input into the ANATES software according to the number of students. The researcher used ANATES software version 4.09 to analyze the difficulty levels and discriminatory power of the TOAFL items. The data presented in this study are in the form of tables and diagrams, which are then interpreted and expressed in percentages.

Result

1. *Difficulty Levels of TOAFL Items*

In addition to being valid and reliable, the quality of a test item is evaluated based on the balance of its discriminatory power. This balance refers to the presence of questions that vary proportionally across easy, moderate, and difficult levels. The difficulty level of a question is assessed based on students' ability to answer it, rather than from the perspective of lecturers as the test designers (Fatimah & Alfath, 2019). Test items considered difficult or easy by lecturers may not necessarily reflect the same level of difficulty for students.

An ideal test question should strike a balance in its difficulty level, ensuring that it is neither too easy nor too difficult so that it can accurately measure students' abilities. Questions that are too easy, where all students can answer correctly, are deemed less effective as they do not encourage critical thinking. Conversely, questions that are too difficult, where no student can provide the correct answer, are also considered suboptimal (Prastika, 2021). Therefore, an ideal test item is one with a balanced difficulty level, typically ranging between 0.15 and 0.85 (Oller dalam Ainin, 2023).

Moreover, an ideal score category is characterized by a distribution of questions across varying levels of difficulty. The category of easy questions should account for approximately 20-30% of the total questions. Moderate questions should comprise 40-60% of the total, while difficult questions should make up 19-20% of the total (R. S. Qomariyah, 2022).

The formula used to measure the difficulty level of a test is as follows

$$\boldsymbol{Test\ Item\ Difficulty\ Level\ (P)} = \frac{\boldsymbol{Number\ of\ Correct\ Answers}}{\boldsymbol{Number\ of\ Test\ Participants}} \times 100\%$$

According to Witherington in his book *Psychological Education*, the difficulty level of test items measuring learning outcomes can be assessed by calculating the difficulty index, which ranges from 0.00 to 1.00. The higher the difficulty index obtained from the calculation, the easier the test item is. Conversely, the lower the difficulty index, the more difficult the test item becomes (Dianova & Anwar, 2024). For instance, a difficulty index of 0.0 indicates that the test item is extremely difficult, whereas an index of 1.0 suggests that the test item is very easy (Solichin, 2017).

Table 1

Index of Test Difficulty Level

Value of P	Categories
0,00	Very difficult
0,00 < P ≤ 0,30	Difficult
0,31 < P ≤ 0,70	Moderate
0,71 < P < 1,00	Easy
1,00	Very easy

Sudijono recommends several follow-up actions after conducting an analysis of the difficulty levels of test items. Referring to Table 1, items categorized as good can be compiled into a question bank for reuse in subsequent tests. Meanwhile, items classified as very difficult can be revised and excluded from future tests, further examined for improvement, or retained for use in highly stringent assessments. Test items considered very easy may be evaluated and excluded from future tests, reviewed for improvement and reused in subsequent assessments, or retained for use in flexible tests (Fitriani, 2021).

Based on the analysis of the TOAFL test items at UNWAHA using ANATES, the difficulty levels of the test items are presented in Table 2.

Table 2

The Difficulty Levels of Test Items of Listening Skill

Questions	**Number of Questions**	**Difficulty Level**	**Categories**
11, 24, 36, 37	4	0,00	Very difficult
6, 8, 9, 14, 16, 17, 19, 20, 23, 26, 27, 29, 30, 31, 32, 47, 48, 50	18	0,00–0,30	Difficult
1, 2, 3, 4, 5, 7, 10, 12, 13, 15, 18, 21, 22, 25, 28, 33, 34, 35, 38, 39, 40, 41, 42, 44, 45, 46, 49	27	0,31-0,70	Moderate
43	1	0,71-1,00	Easy
-	-	1,00	Very easy

Table 3

The Difficulty Levels of Test Items of Writing Skill

Questions	**Number of Questions**	**Difficulty Level**	**Categories**
31, 33, 38, 39	4	0,00	Very difficult
2, 4, 5, 20, 21, 32, 36, 40	8	0,00–0,30	Difficult
1, 3, 6, 7, 8, 9, 10, 11, 16, 17, 19, 22, 23, 24, 26, 27, 28, 29, 30, 34, 37	21	0,31-0,70	Moderate
12, 13, 14, 18, 25, 35	6	0,71-1,00	Easy
15	1	1,00	Very easy

Table 4

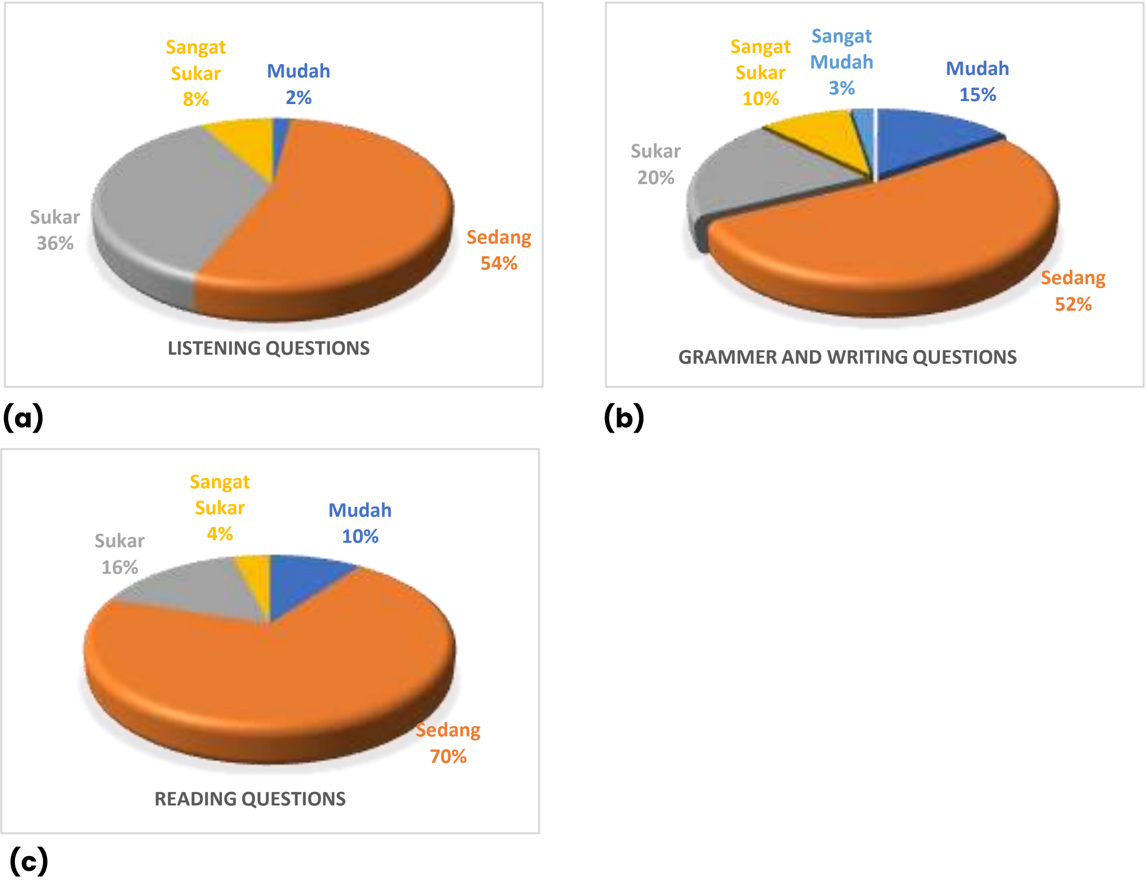
The Difficulty Levels of Test Items of Reading Skill

Questions	**Number of Questions**	**Difficulty Level**	**Categories**
6, 23	2	0,00	Sangat sulit
9, 22, 26, 28, 31, 33	6	0,00–0,30	Sulit
1, 2, 3, 4, 5, 7, 8, 10, 12, 13, 14, 16, 18, 19, 20, 21, 24, 25, 27, 29, 30, 32, 35, 36, 37, 38, 39, 40, 41, 43, 44, 46, 47, 48, 49, 50	37	0,31-0,70	Sedang
11, 15, 34, 42, 45	5	0,71-1,00	Mudah
-	-	1,00	Sangat mudah

Based on Tables 2, 3, and 4, the percentage of item difficulty levels for each skill can be determined by referring to the following diagram:

Figure 1

*The Difficulty Level of TOAFL Test Items Based on Skills **(a)** Listening Questions; **(b)** Grammer and Writing Questions; **(c)** Reading Questions*



The results of the difficulty level analysis for the TOAFL test items in the listening skills section are presented in Figure A. This figure shows that 1 item (2%) falls into the easy category, 27 items (54%) are in the moderate category, 18 items (36%) are in the difficult category, and 4 items (8%) are in the very difficult category, specifically items numbered 11, 24, 36, and 37. Based on the difficulty index, these items are still suitable for use in future tests.

The results of the difficulty level analysis for the TOAFL test items assessing the skills of understanding structure and writing are presented in Figure B. This figure indicates that 1 item (2.5%) is categorized as very easy, and 6 items (15%) fall into the easy category. A total of 21 items (52.5%) are in the moderate category, while 8 items (20%) are classified as difficult, and 4 items (10%) are categorized as very difficult, specifically items numbered 81, 83, 88, and 89. Based on the difficulty index, these items are also suitable for use in future tests.

The results of the difficulty level analysis for the TOAFL test items in the reading skills section are shown in Figure C. This figure reveals that 5 items (10%) are in the easy category, 35 items (70%) fall into the moderate category, 8 items (16%) are in the difficult category, and 2 items (4%) are in the very difficult category, specifically items numbered 96 and 113. Based on the difficulty index, these items remain appropriate for inclusion in future tests.

2. Differentiating power of TOAFL Items

The quality of a test item is not solely measured by its difficulty level but also by its discriminatory power. Discriminatory power refers to a test item's ability to differentiate between high-performing and low-performing students (Loka Son, 2019). According to Djiwandono, a test item is considered good in terms of discriminatory power if a higher number of students from the high-performing group (H) correctly answer the item compared to students from the low-performing group (L) answering the same item correctly (Ahsanuddin, 2016).

$$\boldsymbol{Discriminatory\ Power\ (D)} = \frac{\sum \boldsymbol{Correct\ upper} - \sum \boldsymbol{Correct\ lower}}{\sum \boldsymbol{Group\ (upper\ atau\ lower)}} \times 100\%$$

To calculate the discriminatory power, the number of participants from the high-performing group who answered correctly is subtracted from the number of participants from the low-performing group who answered correctly. The result is then interpreted based on four criteria (Djiwandono dalam Ainin, 2023), as follows: .

Table 5

Discrimination Index

Discrimination Index Range	Criteria
0,00-0,19	Poor
0,20-0,39	Fair
0,40-0,69	Good
0,70-1,00	Excellent

Note: (1) ***Excellent***: Indicates that the test item effectively distinguishes between high-performing and low-performing students. **(2)** ***Good***: Suggests that the test item has a satisfactory level of discriminatory power. **(3)** ***Fair***: Implies that the test item's ability to differentiate between student performance levels is moderate but still usable. **(4)** ***Poor***: Indicates that the test item does not adequately distinguish between high-performing and low-performing students, requiring revision or exclusion from the test

The quality of a test item is directly proportional to its discriminatory power index. A test item is considered high-quality if it has a high index, as it effectively differentiates between high-performing and low-performing students. Conversely, the lower the index, the less effective the test item is in distinguishing students' abilities. **(Nurhalimah et al., 2022)**.

Based on the analysis of the TOAFL test at UNWAHA using ANATES, the discriminatory power is presented in Table 6.:

Table 6

Discriminatory power index of TOAFL test items

Questions	Number of Questions	Indexs of Discriminatory Power	Categories
6, 8, 16, 20, 31 35, 36, 37, 50, 81, 82, 89, 90, 99, 101, 103, 107, 113, 118, 132	20	-	Negative
4, 5, 9, 11, 14, 23, 26, 27, 30, 32, 57, 86, 95, 104, 108, 115, 123, 125	18	0,00–0,19	Low
19, 33, 43, 47, 48, 52, 55, 60, 65, 66, 69, 70, 71, 79, 84, 97, 106, 112, 117, 128, 130, 131, 135	23	0,20-0,39	Medium
1, 7, 10, 12, 13, 17, 21, 22, 24, 25, 28, 29, 34, 39, 40, 41, 42, 45, 46, 49, 51, 53, 54, 56, 58, 59, 63, 64, 68, 72, 75, 77, 83, 85, 87, 88, 91, 92, 93, 94, 96, 98, 100, 102, 105, 114, 116, 119, 120, 121, 122, 124, 126, 129, 133, 134, 137, 139, 140	59	0,40-0,69	High
2, 3, 15, 18, 38, 44, 61, 62, 67, 73, 74, 76, 78, 80, 109, 110, 111, 127, 136, 138	20	0,70-1,00	Very Higher

Based on Table 5, it is observed that in the listening section, there are 30 items with high discriminatory power and 6 items with very high discriminatory power, 5 items with moderate discriminatory power, 10 items with low discriminatory power, and 9 items with negative discriminatory power. In the writing section, there are 16 items with high discriminatory power and 8 items with very high discriminatory power, 11 items with moderate discriminatory power, 3 items with low discriminatory power, and 4 items with negative discriminatory power.

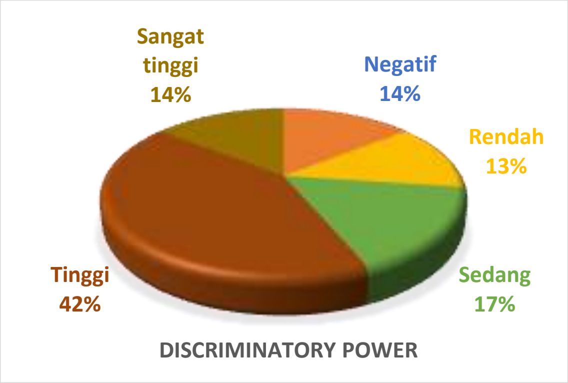
Meanwhile, in the reading section, there are 23 items with high discriminatory power and 6 items with very high discriminatory power, 8 items with moderate discriminatory power, 6 items with low discriminatory power, and 7 items with negative discriminatory power.

When summed across all skill aspects of the TOAFL test, there are 59 items with high discriminatory power (42.1%) and 20 items with very high discriminatory power (14.3%). Meanwhile, 23 items have moderate discriminatory power (16.5%), and test items within these two categories can still be retained. However, 20 items fall into the negative category (14.3%), and 18 items have low discriminatory power (12.8%), making them ineffective in distinguishing between high-performing and low-performing students. These items should be removed or replaced with new ones.

To facilitate understanding, a diagram is presented illustrating the categories of discriminatory power for the analyzed TOAFL test items.

Figure 2

Discriminatory Power of TOAFL test Items



Based on the diagram, it can be observed that the overall percentage of the discriminatory power of the TOAFL test items at UNWAHA, across all language skills, falls into the high category. More than half of the test items, specifically 56.4%, have good discriminatory power. This indicates that the items are effective in distinguishing between high-performing and low-performing students. However, some items still fail to differentiate between high and low-performing students (particularly those with low and negative discriminatory power), and therefore, they need to be revised to improve the overall quality of the test items.

Conclusion

Based on the analysis of the TOAFL test items at Universitas KH. Abdul Wahab Hasbullah, it can be concluded that the difficulty level of the test items is considered adequate, with a balanced proportion across each skill. There are 13 items (9.3%) categorized as easy, 83 items (59.3%) categorized as moderate, and 44 items (31.5%) categorized as difficult. The discriminatory power of the items is also deemed good (56.4%), as more than 50% of the items are able to identify the differences between high-performing and low-performing student groups. Only a small portion of the items has low (17.8%) and negative (26.4%) discriminatory power, which requires revision.

The findings of this study suggest that items with moderate difficulty and moderate or high discriminatory power should be reused in future tests. Meanwhile, items with difficult difficulty levels and moderate or high discriminatory power should be further examined to understand the reasons behind students' difficulties in answering those items. Items with easy difficulty levels and moderate or high discriminatory power can be retained.

Declarations

Author contribution statement

The first author (Bustanil Ilmi Agustin) was responsible for research planning, data collection, and analysis. The second author (Rikha Ikke Nuriani) contributed to the methodology, data processing, and manuscript drafting. The third and fourth authors (Nur Qurrotaa'yun and Baihaqi) provided input and revisions prior to submission.

Funding statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

Data availability statement

The data supporting the findings of this study are stored at the Language and Computer Laboratory of UNWAHA. The data can be accessed by other researchers upon request to the corresponding author.

Declaration of interests statement

The authors declare no financial interests or personal relationships that could influence the outcomes of this research.

Additional information

This study was conducted during the even semester of 2024. The TOAFL test items analyzed were one of three test packages used as a graduation requirement for undergraduate students.

References

Ahsanuddin, M. (2016). ANALISIS HASIL TOAFL MAHASISWA JURUSAN SASTRA ARAB FAKULTAS SASTRA UNIVERSITAS NEGERI MALANG. *Prosiding Konferenasi Nasional Bahasa Arab II (KONASBARA)*. https://prosiding.arab-um.com/index.php/konasbara/article/view/76

Ainin, M. (2016). KESAHIHAN DALAM PENYUSUNAN TES BAHASA ARAB DI MADRASAH/SEKOLAH. *Prosiding Konferenasi Nasional Bahasa Arab II (KONASBARA)*. https://prosiding.arab-um.com/index.php/konasbara/article/view/75

Ainin, M. (2023). *Penilaian Berpikir Tingkat tinggi (HOTS) dalam Pembelajaran Bahasa Arab*. CV Bintang Sejahtera.

B., N. (2020). *Analisis Butir Soal TOAFL pada Jurusan Pendidikan Bahasa Arab Fakultas Tarbiyah dan Keguruan UIN Alauddin Makassar* [Tesis]. https://repositori.uin-alauddin.ac.id/18284/1/NURHAYATI.pdf

Bamualim, M. (2020). Kedudukan Dan Tujuan Evaluasi Pembelajaran Bahasa Arab. *Jurnal Al-Fawa'id : Jurnal Agama dan Bahasa*, *10*(2), 1–10. https://doi.org/10.54214/alfawaid.Vol10.Iss2.141

Fahmi, B., Rizqi, S., & H, N. E. (2022). *ANALISIS BUTIR SOAL BAHASA ARAB SISWA MAS PONDOK PESANTREN ASAALAM KAMPAR RIAU*. *6*(2), 95–105. https://doi.org/DOI: 10.15575/jpba.v5i2. 16193

Fatimah, L. U., & Alfath, K. (2019). ANALISIS KESUKARAN SOAL, DAYA PEMBEDA DAN FUNGSI DISTRAKTOR. *Al-Manar*, *8*(2), 37–64. https://doi.org/10.36668/jal.v8i2.115

Fitriani, N. (2021). ANALISIS TINGKAT KESUKARAN, DAYA PEMBEDA, DAN EFEKTIVITAS PENGECOH SOAL PELATIHAN KEWASPADAAN KEGAWATDARURATAN MATERNAL DAN NEONATAL. *Paedagoria : Jurnal Kajian, Penelitian dan Pengembangan Kependidikan*, *12*(2), 199. https://doi.org/10.31764/paedagoria.v12i2.4956

Halomoan, H., Ibrahim, F. M. A., & Bahruddin, U. (2022). Taḥlīl Ikhtibār Kafā`ah al-Lugah al-'Arabiyyah li an-Nāṭiqīna bigairihā fī Jāmi'ah Sulṭān Syarīf Qāsim al-Islāmiyyah al-Ḥukūmiyyah Riau. *LISANIA: Journal of Arabic Education and Literature*, *6*(1), 74–87. https://doi.org/10.18326/lisania.v6i1.74-87

Hamdun, D., & Islam, N. (2023). Humanistic Approaches in Learning Arabic to Increase Motivation of Students' Learning: Pendekatan Humanistik dalam Pembelajaran Bahasa Arab Upaya Meningkatkan Motivasi Belajar Siswa. *Edulab : Majalah Ilmiah Laboratorium Pendidikan*, *8*(2), 177–193. https://doi.org/10.14421/edulab.2023.82.05

Harahap, P. (2018). ANALISIS SOAL TOAFL IAIN CURUP. *Ihya Al-Arabiyah*, *2*. http://dx.doi.org/10.30821/ihya.v4i2.3089

Ishak, D. M., & Fitriyanti, E. N. (2020). PENGARUH PEMBELAJARAN BAHASA ARAB MAHARAH QIRA'AH UNTUK SISWA MADRASAH ALIYAH TERHADAP PEMAHAMAN BUDAYA ARAB. *Prosiding Seminar Bahasa Arab Mahasiswa IV UM*. https://prosiding.arab-um.com/index.php/semnasbama/article/view/579/532

Loka Son, A. (2019). INSTRUMENTASI KEMAMPUAN PEMECAHAN MASALAH MATEMATIS: ANALISIS RELIABILITAS, VALIDITAS, TINGKAT KESUKARAN DAN DAYA BEDA BUTIR SOAL. *Gema Wiralodra*, *10*(1), 41–52. https://doi.org/10.31943/gemawiralodra.v10i1.8

Muhammad Lukman Arifianto, Moh. Ainin, Irhamni, Mohammad Ahsanuddin, Khoirin Nikmah, Mohammad Sofi Anwar, & Nurul Fitria. (2021). *Evaluasi Pembelajaran Bahasa Arab dan Pengembangan Tes Interaktif*. Tonggak Media.

Muhimmatul Choiroh. (2021). EVALUASI PEMBELAJARAN BAHASA ARAB BERBASIS MEDIA E-LEARNING. *Jurnal Naskhi: Jurnal Kajian Pendidikan dan Bahasa Arab*, *3*(1), 41–47. https://doi.org/10.47435/naskhi.v3i1.554

Nurhalimah, S., Hidayati, Y., Rosidi, I., & Hadi, W. P. (2022). HUBUNGAN ANTARA VALIDITAS ITEM DENGAN DAYA PEMBEDA DAN TINGKAT KESUKARAN SOAL PILIHAN GANDA PAS. *Natural Science Education Research*, *4*(3), 249–257. https://doi.org/10.21107/nser.v4i3.8682

Pradita, E., & Megawanti, P. (2023). Analisis Tingkat Kesukaran, Daya Pembeda, dan Fungsi Distraktor PTS Matematika SMPN Jakarta. *Himpunan: Jurnal Ilmiah Mahasiswa Pendidikan Matematika*. https://doi.org/10.36706/jptm.v1i2.7410

Pranata, A. (2022). *PENGEMBANGAN INSTRUMEN TES KETERAMPILAN MENYIMAK DALAM PEMBELAJARAN BAHASA ARAB BAGI PESERTA DIDIK*.

Prastika, Y. D. (2021). PENGARUH VALIDITAS, RELIABILITAS DAN TINGKAT KESUKARAN TERHADAP KUALITAS BUTIR SOAL EKONOMI MENGGUNAKAN SOFTWARE ANATES DI SMKN 3 BANGKALAN. *STKIP PGRI Bangkalan*. http://repo.stkippgri-bkl.ac.id/id/eprint/1092

Qodri, M. (2020). *Problematika Pembelajaran TOAFL Pada Mahasiswa Fakultas Ilmu Tarbiyah dan Keguruan IAIN Sulthan Thaha Saifuddin Jambi*. *1*(1). https://doi.org/10.36915/la.v1i1.1

Qomariyah, L. (2022). Analisis Tingkat Kesukaran dan Daya Pembeda Butir Soal TOAFL Universitas Hasyim Asy'ari Tebuireng Jombang. *Lisanan Arabiya: Jurnal Pendidikan Bahasa Arab*, *6*(1), 1–18. https://doi.org/10.32699/liar.v6i1.2549

Qomariyah, L., & Niswah, I. (2021). ANALISIS KUALITAS BUTIR SOAL TOAFL UNHASY TEBUIRENG. *Prosiding Seminar Nasional SAINSTEKNOPAK Ke 5 LPPM UNHASY*.

Qomariyah, R. S. (2022). *Analisis Tingkat Kesukaran Dan Daya Pembeda Pada Butir Soal Pilihan Ganda Mata Pelajaran Bahasa Indonesia Kelas V Semester 1 SDN Kedungdalem 2*. *1*(2).

Rathomi, A. (2020). *MAHARAH KITABAH DALAM PEMBELAJARAN BAHASA ARAB*. https://doi.org/10.37567/ti.v1i1

Salam, M. Y., Zela, A. F., Helviza, D., & Ikhlas, Z. (2023). *PENGEMBANGAN PANDUAN TES BAHASA ARAB BERBASIS TOAFL DI MADRASAH ALIYAH NEGERI 2 PADANG PANJANG MENGGUNAKAN APLIKASI LECTORA INSPIRE, QUIZIZZ, DAN KAHOOT*. *12*(2).

Solichin, M. (2017). ANALISIS DAYA BEDA SOAL, TARAF KESUKARAN, VALIDITAS BUTIR TES, INTERPRETASI HASIL TES DAN VALIDITAS RAMALAN DALAM EVALUASI PENDIDIKAN. *Dirasat: Jurnal Manajemen dan Pendidikan Islam*, *2*. https://doi.org/10.26594/dirasat.v2i2.879

Susiawati, I., Mardani, D., & Indramayu, I. A.-A. (2022). Bahasa Arab Bagi Muslim Indonesia antara Identitas dan Cinta pada Agama. *Jurnal Pendidikan dan Konseling*, *4*. https://doi.org/10.31004/jpdk.v4i5.5432

Taufik, T., Syifaanddini, S., Rochmahtika, A. S., Qiyamullaily, A. J., Nazhifah, A. I., & Azharie, H. J. (2023). Media Pembelajaran Busuu Dalam Pembelajaran Bahasa Arab Untuk Pemula. *Al Mi'yar: Jurnal Ilmiah Pembelajaran Bahasa Arab Dan Kebahasaaraban*, *6*(2), 749. https://doi.org/10.35931/am.v6i2.2321

Utami, R. W. (2018). *Tahlil Bunudi As'ilati Ikhtibaraat al-Lughah al-Arabiyyah Li Ghairi an-Natiqina Biha 'Ala Dhou'i Nadhoriyyah Tashnif Bloom (Dirasah Washfiyyah Tahliliyyah Fi al-Jami'ah al-islamiyyah al-Hukumiyyah Ponorogo)*. Universitas Islam Negeri Maulana Malik Ibrahim Malang. http://etheses.uin-malang.ac.id/12295/1/16720015.pdf

Wulandari, N. (2023). Tahlil Ikhtibarat Al-Arabiyah ka Lughah Ajnabiyah. *An Nabighoh*, *25*(1), 77. https://doi.org/10.32332/an-nabighoh.v25i1.7001