Jurnal Penelitian dan Evaluasi Pendidikan
Volume 29, No. 2, December 2025 (271-284)
Online: http://journal.uny.ac.id/index.php/jpep

A multidimensional item response theory approach in the item analysis of Arabic language tests in madrasah aliyah

Joko Subando[1*]; Dudi Budi Astoko[1]; Lailla Hidayatul Amin[1]; Budiharjo[2]; Rahimah Embong[3]
[1]Institut Islam Mambaul Ulum Surakarta, Indonesia
[2]Sekolah Tinggi Islam Al Mukmin Surakarta, Indonesia
[3]Universiti Sultan Zainal Abidin Trengganu, Malaysia
[*]Corresponding author. E-mail: jokosubando@yahoo.co.id

ARTICLE INFO

Article History
Submitted:
28 October 2025
Revised:
3 December 2025
Accepted:
24 December 2025

Keywords
Arabic language; item analysis; item response theory; multidimensional item response theory

Scan Me:

ABSTRACT

This study evaluates the quality of Arabic test items in madrasah assessments using a quantitative approach based on Multidimensional Item Response Theory (MIRT). The sample comprised 321 twelfth-grade students from MAN 1 Surakarta, purposively selected because the institution implements systematic and independent assessments. Data were obtained from student responses to the final Arabic examination in the 2022/2023 academic year. Exploratory Factor Analysis (EFA) was first conducted to identify the dimensional structure of the test, using the criteria KMO > 0.60 and a significant Bartlett's Test of Sphericity ($p < 0.05$). Factor extraction was determined by eigenvalues > 1 and supported by scree plot inspection. Model fit was subsequently examined using a MIRT 2-parameter logistic (2PL) model in R, with evaluation indicators RMSEA < 0.06, CFI > 0.90, and TLI > 0.90. Item parameters included discrimination (d) and difficulty (b), where discrimination was classified as: < 0.00 (unacceptable); 0.00–0.34 (very low); 0.35–0.64 (low); 0.65–1.34 (moderate); ≥ 1.35 (high). Findings show substantial variability in item performance. Most items demonstrated acceptable discrimination; however, 16 items had negative discrimination, indicating weaknesses in content representation and item construction. A few items (items 1, 3, 7, 10, and 22) showed high discrimination and are highly informative. Difficulty levels were dominated by easy items, limiting the test's ability to distinguish medium- to high-ability examinees. The study recommends revising misfitting items, adding items with moderate difficulty and d > 0.65, and enhancing validity through Confirmatory Factor Analysis and bias detection using DIF analysis.

To cite this article (in APA style):
Subando, J., Astoko, D. B., Amin, L. H., Budiharjo, & Rahimah, E. (2025). A multidimensional item response theory approach in the item analysis of Arabic language tests in madrasah aliyah. *Jurnal Penelitian dan Evaluasi Pendidikan*, *29*(2), 271-284. https://doi.org/10.21831/pep.v29i2.90877

INTRODUCTION

In the field of education, the quality of assessment instruments is a critical component that cannot be overlooked (Ningsih et al., 2025). Measurement instruments such as tests or item packages serve as tools to assess students' learning achievement, evaluate instructional effectiveness, and support decision-making at various levels, from classroom teachers to education policymakers (Engida et al., 2024). A well-constructed instrument not only measures what it is intended to measure, but does so accurately, fairly, and objectively (Terry & Nguyen, 2024). Therefore, the development and quality analysis of test packages is a crucial step in ensuring the integrity of educational assessments.

A high-quality test package is one that can yield valid and reliable data (Asadizanjani et al., 2025). The items within the package must have an appropriate level of difficulty, good

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i2.90877

discriminatory power, and be free from cultural, gender, or social group biases (Belenguer, 2022). A sound instrument should also be based on clearly defined ability constructs and measured systematically using robust measurement theories (Freed et al., 2022). In this context, modern measurement theories such as Item Response Theory (IRT) have become central references in the development and analysis of test items.

IRT offers a different approach from Classical Test Theory (CTT). While CTT item analysis heavily depends on the characteristics of the test-taker sample, IRT provides models that allow for the estimation of item parameters (such as difficulty and discrimination) relatively independently of the sample (Ayanwale et al., 2022). This makes IRT advantageous for developing test items that are more stable and generalizable. As a result, IRT is widely used in large-scale testing programs, such as national examinations, university entrance exams, and computer-based assessments (Oladele & Ndlovu, 2021).

However, IRT relies on several assumptions to produce valid results. One of its core assumptions is unidimensionality, the idea that all items in a test measure a single underlying trait or ability (Wilson, 2023). This implies that test-takers are assumed to possess one latent ability that drives their responses to all items. While this assumption simplifies analysis and interpretation, it can also become a significant limitation when applied to tests that are inherently multidimensional.

In practice, many test packages cannot be considered unidimensional. This is especially evident in complex subjects that encompass multiple cognitive skills, such as language learning. Arabic language tests, for example, are designed to assess various competencies including reading comprehension (*qirā'ah*), grammatical structure (*nahwu and sharaf*), and vocabulary mastery (*mufradāt*) (Nury et al., 2025; Zakkiyah et al., 2024). These three aspects have distinct characteristics and tend to form separate dimensions of ability. Thus, applying a unidimensional IRT model to Arabic test items risks producing inaccurate analyses by neglecting the complexity of the construct being measured.

The multidimensional nature of Arabic language test items is not only an academic concern but also reflects the language learning's psychological and cognitive realities (Alhamami, 2025; Zeinoun et al., 2022). Learning a language involves integrating various linguistic and cognitive components. When a student reads an Arabic text, they must apply grammatical understanding, recall relevant vocabulary, and interpret meaning within a broader context (Saepudin et al., 2024). Therefore, the construction and analysis of Arabic language assessments require approaches that can more accurately and comprehensively capture this complexity.

To address this challenge, Multidimensional Item Response Theory (MIRT) emerges as a more suitable solution. MIRT is an extension of IRT that removes the restriction of unidimensionality (Ackerman & Ma, 2024; Jewsbury & van Rijn, 2020). In MIRT, each item can be associated with multiple ability dimensions (Lee et al., 2020). This allows for more realistic analysis of item and test data, especially for subjects that are conceptually and practically multidimensional, such as Arabic.

Using the MIRT approach, item analysis can identify the extent to which each item contributes to a particular dimension of the measured ability (Mardapi, 2020). For instance, one item may primarily assess grammatical knowledge, while another may focus more on contextual understanding or synthesising information from reading passages. By employing MIRT, test developers can gain deeper insights into the test package's dimensional structure and revise inconsistent or overly dominant items in a single dimension (Ntumi, 2025). It will facilitate the development of balanced instruments that proportionally reflect all aspects of students' abilities.

In Indonesia, the application of the MIRT approach in item analysis remains relatively new and underexplored, particularly in the context of madrasah education. Based on a review of the literature and previous studies, to date, no research has specifically analyzed the quality of Arabic language test items using a multidimensional approach in madrasah settings, especially at MAN 1 Surakarta. Most existing studies still rely on classical or unidimensional IRT

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i2.90877

approaches. For example, Mahmudi et al. (2023) analyzed Arabic test items using Classical Test Theory, as by Kadir et al. (2024) and Jundi (2023). Meanwhile, Ramadhan and Subando (2025) conducted item analysis for Fiqh test items using unidimensional IRT, a method also used by Madi and Clinton (2015) and Al-Qerem et al. (2025). Given Arabic as a subject's inherently multidimensional nature, classical test theory is insufficient to capture the complexity of the constructs being measured. This constitutes the main rationale for conducting the present study.

This study aims to analyze the quality of Arabic language test items used in madrasah assessments at MAN 1 Surakarta using the MIRT approach. It evaluates item parameters such as discrimination and difficulty levels across each involved dimension. In addition, it seeks to identify the multidimensional structure of the test package and provide recommendations for revising items that are misaligned or require adjustment. The findings of this study are expected to serve as a foundation for developing improved assessment instruments in the future.

This topic is crucial for several reasons. (1) This study contributes to the advancement of educational measurement, particularly in applying MIRT to Arabic language assessments. (2) It enriches the currently limited body of literature on multidimensional test item analysis in madrasah contexts. (3) It provides practical benefits for teachers, curriculum developers, and item writers in designing more accurate, representative, and balanced assessment instruments.

RESEARCH METHOD

This study is an evaluative study aimed at assessing the quality of Arabic test items in the madrasah assessment using a quantitative approach. The evaluation addresses construct validity and item characteristics by employing a Multidimensional Item Response Theory (MIRT) two-parameter logistic (2PL) model. The study focuses on evaluating the extent to which the Arabic test items can measure dimensions of students' abilities accurately and representatively.

The research sample consisted of 321 twelfth-grade (Grade XII) students from *Madrasah Aliyah Negeri* (MAN) 1 Surakarta. MAN 1 Surakarta was purposively selected because it is one of the leading madrasahs in the Surakarta area and has implemented madrasah assessments independently and systematically. Data were collected from the madrasah assessment examinations administered in the 2022/2023 academic year, specifically for the Arabic subject. Student responses, coded dichotomously (correct = 1, incorrect = 0), were analyzed to identify the dimensional structure and to evaluate item quality.

Data analysis procedures: (1) Exploratory Factor Analysis to determine the number of factors or dimensions underlying students' responses to the items. An exploratory factor analysis was conducted using SPSS (Puia et al., 2025). This step aimed to capture the test's latent structure and ensure that a multidimensional approach was warranted. Evaluation criteria at this stage were: (a) data adequacy with a Kaiser–Meyer–Olkin (KMO) value > 0.60, and a significant Bartlett's Test of Sphericity (BTS) ($p < 0.05$). The number of factors was determined based on eigenvalues > 1, with a scree plot used to support the decision on the number of factors (Stalikas et al., 2018; Watkins, 2018). (2) MIRT Model Fit Analysis, after the factors were identified, data–model fit was examined using an MIRT model implemented in R. The MIRT model used was the two-parameter logistic (2PL), which accounts for the discrimination and difficulty parameters for each item on each dimension (Jewsbury & van Rijn, 2020). Model-fit evaluation criteria were: (a) Root Mean Square Error of Approximation (RMSEA) < 0.06; (b) Comparative Fit Index (CFI) > 0.90; and (c) Tucker–Lewis Index (TLI) > 0.90. (3) Item Parameter Analysis, Item parameters were estimated under the MIRT 2PL model. The parameters analyzed included the discrimination parameter (d), categorized as follows: $d < 0.00$ = not of acceptable quality; 0.00–0.34 = very low; 0.35–0.64 = low; 0.65–1.34 = moderate; $d \geq 1.35$ = high (Pardede et al., 2023). The results of this analysis were used to evaluate the quality of each item based on how well it discriminated among students with different ability levels on a given dimension.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i2.90877

FINDINGS AND DISCUSSION

Findings

Before analyzing the Arabic test items using Item Response Theory (IRT), it is essential to examine whether the students' response data satisfy the underlying assumptions. Conducting IRT when the assumptions are not met may generate biased parameter estimates and lead to incorrect conclusions. The fundamental assumptions in IRT include unidimensionality, local independence, and parameter invariance. To evaluate unidimensionality, Exploratory Factor Analysis (EFA) was conducted using SPSS. The results showed that the dataset met the minimum adequacy requirements, indicated by a Kaiser–Meyer–Olkin (KMO) value of 0.914 (> 0.60) and a significant Bartlett's test of sphericity with p = 0.000 (< 0.05), confirming that factor analysis was appropriate for the data (see Table 1).

Table 1. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		**.914**
Bartlett's Test of Sphericity	Approx. Chi-Square	3.029E3
	Df	666
	Sig.	.000

Following data adequacy confirmation, the factor analysis revealed the presence of 10 factors with eigenvalues greater than 1. The first factor had an eigenvalue of 8.639 and accounted for 23.348% of the total variance, demonstrating that it did not dominate the measurement structure. In standard IRT analysis, a dominant factor typically explains at least 40% of the variance, providing strong evidence of unidimensionality. Since this requirement was not fulfilled, the findings suggest that the test measures multiple underlying ability dimensions rather than a single construct (see Table 2). Therefore, relying on unidimensional IRT would risk misrepresenting item characteristics and student ability profiles, prompting the use of Multidimensional IRT (MIRT) as a more appropriate analytical approach.

Table 2. Nilai Eigen Value

Component	**Initial Eigenvalues**		
	Total	**% of Variance**	**Cumulative %**
1	8.639	23.348	23.348
2	1.721	4.652	28.001
3	1.513	4.089	32.089
4	1.366	3.691	35.780
5	1.236	3.341	39.121
6	1.220	3.298	42.419
7	1.160	3.134	45.553
8	1.051	2.841	48.394
9	1.026	2.772	51.166
10	1.006	2.719	53.886

In the context of MIRT, factor loadings represent the strength of the relationship between an item and a particular latent factor. Items with loadings above 0.30 are considered meaningful contributors to the construct being measured. For instance, Item 1 shows the highest loading on Factor 6 (F6 = 0.524), indicating that it primarily measures that dimension. Likewise, Item 24 exhibits a very high loading on Factor 2 (F2 = 0.952), suggesting a strong association with that factor (see Table 3). These values provide evidence of how items cluster into different ability domains relevant to Arabic language competency.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i2.90877

Table 3. MIRT Factor Loadings

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	h2
Item 1	-0.03459	0.2067	-0.00179	0.065116	-0.143636	0.52381	0.09623	0.16845	0.15219	0.0193	0.687
Item 2	0.05335	0.0957	0.02907	0.084819	0.107242	0.63664	0.06132	-0.03239	-0.02890	0.0387	0.564
Item 3	0.22400	0.1248	0.23336	0.232349	0.007959	0.32732	-0.01261	0.07942	-0.10822	0.2502	0.824
Item 7	0.34794	-0.1632	-0.02319	0.267484	-0.039206	0.20120	0.13302	0.06342	0.37239	0.0479	0.781
Item 8	0.23119	0.1281	0.12209	0.167649	-0.072797	-0.08170	0.09708	0.14599	0.17419	0.1853	0.535
Item 9	0.14659	-0.0539	0.17966	0.055241	-0.025412	0.04832	-0.26295	0.11257	0.16273	0.4876	0.624
Item 10	0.15552	0.1693	-0.00585	0.314671	-0.157223	0.10956	-0.12093	-0.01636	0.37496	0.1884	0.855
Item 11	-0.00858	-0.0718	-0.02936	0.203639	0.131539	0.05087	0.02402	0.25363	0.37501	0.3100	0.654
Item 12	-0.05776	0.0365	-0.00485	0.374089	0.094380	-0.00348	0.29957	0.19029	0.02248	0.0132	0.345
Item 13	0.00404	-0.0124	0.41708	-0.150984	-0.097966	-0.03514	0.31030	0.23230	0.13484	-0.3505	0.517
Item 14	-0.02993	-0.0221	0.03228	0.696859	0.032723	0.02417	-0.06904	0.03651	0.07037	-0.0239	0.531
Item 15	0.47416	0.0427	0.26059	0.185450	0.116966	-0.04390	0.12423	0.11355	-0.14467	0.2710	0.735
Item 16	0.02054	0.1900	-0.01825	-0.038625	-0.095770	-0.02942	0.01635	0.17889	0.31200	0.4449	0.624
Item 17	0.21059	-0.0476	0.04003	-0.075557	0.088135	0.37906	0.29906	0.01432	0.22704	0.2416	0.674
Item 18	-0.19679	-0.1347	0.30356	0.191516	0.050868	-0.05961	0.23731	-0.10397	0.06515	0.1048	0.222
Item 19	0.01204	0.0603	0.13022	0.223675	-0.239999	-0.16026	0.31969	-0.01670	-0.02650	0.1307	0.311
Item 20	0.15523	0.3074	0.47812	0.238326	-0.062219	0.08020	-0.12629	0.09016	0.02305	-0.1024	0.732
Item 21	0.18519	0.1580	0.48182	-0.024339	0.062321	0.15618	0.05189	0.07288	0.16289	0.2093	0.799
Item 22	0.01146	-0.0839	0.02056	0.001953	0.045887	-0.01518	0.00815	0.88800	-0.06797	-0.0254	0.734
Item 23	-0.11277	0.0282	0.44798	0.193317	-0.121541	0.18940	-0.08508	0.11805	0.05592	0.1306	0.534
Item 24	-0.01598	0.9524	0.04746	-0.032872	0.017283	0.00980	-0.05928	-0.05938	0.04650	0..1346	0.971
Item 25	0.53354	-0.0749	0.19251	-0.003200	-0.242179	0.09729	0.05166	-0.00129	0.26429	0.0378	0.715
Item 26	-0.00792	0.0875	0.00248	0.041347	-0.000313	0.06184	0.08580	-0.00169	-0.00504	0.8773	0.920
Item 27	0.00920	0.0778	0.17320	0.145501	0.077306	0.04815	0.01657	-0.12283	0.57868	-0.0804	0.466
Item 28	-0.13165	-0.0266	0.14810	0.000208	-0.097059	0.46544	-0.06198	0.09588	0.17916	0.0388	0.376
Item 29	-0.10049	0.3249	-0.02504	0.053667	-0.072925	-0.44573	0.16893	0.21343	0.20529	0.0591	0.374
Item 30	-0.03651	0.0811	-0.03193	0.095884	-0.338458	0.12740	-0.23804	0.26973	0.12758	0.1717	0.478
Item 31	0.03811	0.1113	-0.01802	0.085828	-0.338553	0.08178	0.41654	0.02122	-0.05130	0.2733	0.536
Item 32	0.03808	0.0649	-0.02431	-0.137270	0.043997	0.09604	0.57171	0.09400	0.03354	0.0827	0.421
Item 33	0.21216	0.2281	-0.27577	0.218786	-0.245646	0.19378	-0.04673	0.18894	0.03429	0.0663	0.620
Item 34	0.06789	0.1282	-0.03392	0.120711	0.573642	0.11169	-0.01725	0.23367	0.13826	0.1209	0.604
Item 35	0.62239	0.2414	-0.05186	-0.022388	0.156338	0.05940	0.07949	0.13377	-0.00729	-0.0990	0.622
Item 36	0.12712	0.4155	0.04700	0.108986	0.064305	0.19723	0.16916	0.07268	-0.03810	-0.0593	0.437
Item 37	-0.10106	0.0638	0.19926	-0.009559	0.073222	0.13849	0.16916	0.06941	-0.06122	-0.0916	0.112
Item 38	0.36317	0.1100	-0.09856	0.145471	-0.081305	0.18999	-0.07042	0.07668	0.07040	0.0799	0.499
Item 39	0.12161	0.7226	-0.01596	0.096421	-0.019604	0.04690	0.19802	0.01915	-0.07675	-0.1447	0.696
Item 40	0.32985	0.2425	-0.09637	0.128619	-0.075070	0.15247	-0.03353	0.05284	0.03604	0.0344	0.457

Further evaluation using communality (h^2) values supports the test's multidimensional interpretation. An item with h^2 closer to 1 indicates a high level of variance explained by the factor structure. For example, Item 24 ($h^2 = 0.971$) is considered an excellent item, while Items 18 ($h^2 = 0.222$) and 37 ($h^2 = 0.112$) demonstrate poor item quality and potentially inconsistent behavior. This suggests the need to revise or replace items with low communality, as they contribute less to the measurement objectives.

Table 4. Inter-Factor Correlation

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
F1	1									
F2	0.371	1								
F3	0.237	0.156	1							
F4	0.388	0.356	0.244	1						
F5	-0.017	-0.097	-0.061	-0.108	1					
F6	0.477	0.286	0.209	0.388	-0.052	1				
F7	0.138	0.137	0.173	0.005	-0.038	0.054	1			
F8	0.342	0.247	0.232	0.379	-0.006	0.250	0.203	1		
F9	0.270	0.186	0.262	0.363	-0.171	0.336	0.085	0.326	1	
F10	0.287	0.319	0.137	0.421	-0.086	0.366	0.067	0.254	0.374	1

The pattern of cross-loadings further reinforces multidimensionality. Several items demonstrate significant associations with more than one factor, such as Items 7, 10, 11, 20, 29, and 31. Additionally, some items, including Items 8, 33, and 37, failed to load significantly on any factor, indicating weak construct alignment. Classification of items across factors reveals diverse skill components measured by the test, including vocabulary, grammar, comprehension, and the contextual use of Arabic. The existence of multiple correlated factors (e.g., correlations

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i2.90877

between F1 and F6 = 0.477, F4 and F10 = 0.421, and F6 and F10 = 0.366) also confirms that Arabic language proficiency in this test is composed of interconnected dimensions (Table 4).

To evaluate the items' suitability for MIRT analysis, model-fit indices were assessed for the MIRT 2PL model. The results demonstrated a strong model fit: RMSEA = 0.0226 (< 0.06), CFI = 0.9898 (> 0.90), TLI = 0.9885 (> 0.90), and SMSR = 0.0472 (< 0.08). The p-value was not used due to its high sensitivity to sample size, which can lead to misleading interpretations in large datasets. Since all major fit criteria met acceptable thresholds, the test items were confirmed to be appropriate for evaluation using the MIRT 2PL model (see Table 5). This ensures that further item parameter estimation will produce reliable and interpretable results.

Table 5. Model Fit Test with MIRT 2PL

Indicator	Value	Cutoff	Interpretation
p-value M2	0.0028	> 0.05	The model does not fit perfectly
RMSEA	0.0226	< 0.06	very good
CFI	0.9898	> 0.90	very good
TLI	0.9885	> 0.90	very good
SRMSR	0.0472	< 0.08	very good

Table 6. Item Discrimination Index

	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	d	g	u	b
butir 1	-2.101	-0.026	-0.995	0.072	0.002	0.725	-0.222	-0.295	-0.534	-0.120	1.296	-999	999	NA
butir 2	-1.468	-0.178	-0.597	0.665	-0.457	0.442	-0.498	-0.090	-0.016	-0.305	-0.899	-999	999	NA
butir 3	-3.194	-0.307	-1.263	0.565	-0.497	0.505	-0.018	0.475	0.527	0.617	1.512	-999	999	NA
butir 7	-2.546	-1.076	-1.374	-0.308	0.375	-0.213	-0.046	0.471	-0.222	-0.523	2.843	-999	999	NA
butir 8	-1.508	-0.005	-0.732	-0.400	0.187	-0.130	0.152	0.400	-0.032	0.361	0.662	-999	999	NA
butir 9	-1.448	-0.960	-0.543	-0.333	-0.456	0.293	0.069	0.583	-0.165	0.851	0.624	-999	999	NA
butir 10	-3.497	-0.852	-0.779	-1.170	-0.218	0.881	0.432	0.937	-0.514	0.005	2.223	-999	999	NA
butir 11	-1.458	-0.722	-1.236	-0.904	-0.509	-0.165	-0.114	0.213	-0.288	0.222	1.792	-999	999	NA
butir 12	-0.681	0.245	-0.828	-0.378	-0.046	-0.172	0.041	-0.083	0.349	-0.113	-0.351	-999	999	NA
butir 13	0.066	0.585	-1.116	0.467	0.825	-0.351	0.499	0.125	-0.470	0.025	-2.445	-999	999	NA
butir 14	-1.121	-0.327	-0.601	-0.523	-0.463	0.278	0.701	0.201	0.586	-0.343	-0.593	-999	999	NA
butir 15	-2.193	-0.010	-1.008	0.375	-0.016	-0.686	-0.026	0.658	0.785	0.752	0.731	-999	999	NA
butir 16	-1.536	-0.204	-0.661	-0.953	-0.069	0.206	-0.327	0.326	-0.531	0.726	1.275	-999	999	NA
butir 17	-1.724	-0.365	-1.247	0.227	0.131	-0.086	-1.018	0.338	-0.322	-0.036	0.048	-999	999	NA
butir 18	0.156	0.079	-0.712	-0.102	0.066	0.162	0.016	0.460	0.193	0.028	-0.656	-999	999	NA
butir 19	-0.468	0.343	-0.502	-0.382	0.570	0.211	0.033	0.270	0.304	0.186	-0.039	-999	999	NA
butir 20	-2.059	0.431	-0.930	0.531	-0.174	0.324	1.204	0.769	-0.066	0.411	0.854	-999	999	NA
butir 21	-2.372	0.059	-1.753	0.583	-0.159	0.066	0.113	1.227	-0.456	0.854	0.655	-999	999	NA
butir 22	-0.975	-0.151	-1.524	-0.274	-0.091	-1.083	0.775	-1.428	-0.161	0.874	1.846	-999	999	NA
butir 23	-1.012	-0.194	-1.038	0.121	-0.125	0.654	0.557	0.453	-0.083	0.446	-1.220	-999	999	NA
butir 24	-6.934	5.374	1.144	-1.189	-2.287	1.567	-0.313	1.569	-1.529	1.937	-6.647	-999	999	NA
butir 25	-2.095	-0.662	-0.727	0.321	0.994	-0.093	0.193	0.756	-0.429	0.115	-0.008	-999	999	NA
butir 26	-3.387	-0.857	-1.393	-1.852	-0.952	1.203	-2.381	1.164	0.694	2.471	-3.152	-999	999	NA
butir 27	-0.813	-0.190	-0.706	-0.282	-0.129	0.080	0.236	0.784	-0.644	-0.384	-0.432	-999	999	NA
butir 28	-0.769	-0.392	-0.655	0.196	-0.120	0.573	0.011	-0.006	-0.433	-0.025	0.238	-999	999	NA
butir 29	-0.255	0.670	-0.274	-0.901	0.214	-0.261	0.230	0.095	-0.202	0.337	-0.920	-999	999	NA
butir 30	-1.113	-0.468	-0.204	-0.515	0.069	0.618	0.408	-0.229	-0.340	0.417	0.044	-999	999	NA
butir 31	-1.094	0.366	-0.645	-0.447	0.850	0.494	-0.530	0.037	0.228	0.306	-0.139	-999	999	NA
butir 32	-0.414	0.564	-0.814	-0.021	0.503	-0.326	-0.767	-0.084	-0.025	0.000	-2.136	-999	999	NA
butir 33	-2.026	-0.154	0.024	-0.495	0.138	0.321	0.062	-0.482	0.000	0.000	-0.140	-999	999	NA
butir 34	-1.092	-0.062	-0.763	-0.028	-1.331	-0.890	-0.268	0.000	0.000	0.000	-1.425	-999	999	NA
butir 35	-1.863	0.160	0.060	0.571	-0.079	-0.971	0.000	0.000	0.000	0.000	-0.659	-999	999	NA
butir 36	-1.328	0.570	-0.277	0.193	-0.223	0.000	0.000	0.000	0.000	0.000	-1.152	-999	999	NA
butir 37	-0.072	0.297	-0.446	0.268	0.000	0.000	0.000	0.000	0.000	0.000	-1.052	-999	999	NA
butir 38	-1.663	-0.333	-0.100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.433	-999	999	NA
butir 39	-1.918	1.722	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-1.900	-999	999	NA
butir 40	-1.560	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	NA	0	1	-0.343

Note: *Butir* = item

MIRT was used to estimate the discrimination parameters (a1 through a10) and the intercept/logistic parameter (d) for 40 items under a 10-factor model. The discrimination parameter (commonly denoted "a") indicates an item's ability to distinguish between students with higher and lower ability. The higher an item's discrimination value, the better it differentiates students by ability level.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i2.90877

The observed discrimination values ranged from −6.934 to +5.374, with most items concentrated between −2.5 and +2.5. Items displaying extreme discrimination values, such as Items 24 and 26, initially appear highly informative, but such values frequently indicate model overfitting or instability in the parameter estimation process. These items may not align well with the latent trait structure or may contain ambiguous distractors. Consequently, additional fit diagnostics and content review are required to ensure that the discrimination indices accurately reflect students' Arabic language proficiency rather than measurement error (see Table 6).

Item 1 has the highest discrimination on dimension a1 at −2.101 and on a6 at 0.725. This means the item is most sensitive to changes in examinee ability on the first dimension; however, the negative direction suggests a possible reversed interpretation or miscoding. Item 3 shows high discrimination on dimensions a1 (−3.194) and a9 (0.527), indicating that it measures more than one dimension, particularly the first and ninth. Item 13 shows positive discrimination on dimensions a2 (0.585) and a5 (0.825), with an additional contribution to a7 (0.499), indicating moderate multidimensionality with a tendency toward a5 as the dominant dimension.

In general, the discrimination parameter (a) in these data indicates that most items are effective in measuring examinee ability on one or more dimensions. Dimension a1 appears dominant across many items. Several items with very large or very small values require further scrutiny because they may be misfitting or behaving abnormally. Meanwhile, items with discrimination values of zero or near zero likely provide little meaningful information for measurement and can be considered for revision or removal from the instrument.

Based on the classification of item discrimination values, there is variation in item quality in distinguishing examinee ability. Sixteen items are in the poor category (d < 0.00): Items 12, 13, 14, 18, 19, 23, 24, 26, 32, 33, 34, 35, 36, 37, 39, and 40. Negative discrimination values for these items indicate anomalies in item functioning, whereby high-ability examinees have a lower probability of answering correctly than low-ability examinees. This contradicts fundamental measurement principles and should be followed up with evaluation, revision, or even item elimination. Only one item (Item 17) is in the weak category (0.00–0.34). This item has very limited ability to discriminate among examinees and thus contributes minimally to overall instrument quality. Three items (Items 9, 28, and 30) are categorized as moderate (0.35–0.64), meaning they are reasonably usable in the instrument, though not optimal; they can still provide useful information, especially for examinees in the mid ability range. Next, five items are in the good category (0.65–1.34) (Items 8, 15, 16, 20, and 21). These items show sufficiently sharp and effective discrimination in distinguishing examinees with differing ability levels. Finally, five other items are classified as very good (≥ 1.35): Items 1, 3, 7, 10, and 22. These items are highly effective in differentiating examinees across ability levels and are key components of the instrument's overall quality. Overall, although several items are very good, the predominance of poor or negative items is a major concern that requires prompt evaluation (see Table 7).

Table 7. Classification of Item Discrimination Values

Category	Range of Values	Number of Items	Example Items
Poor/Not acceptable	a < 0	16	12, 13, 14, 18, 19, 23, 24, 26, 32, 33, 34, 35, 36, 37, 39, 40
Very Low	0.00 – 0.34	1	17
Moderate/Fair	0.35 – 0.64	3	9, 28, 30
Good	0.65 – 1.34	5	8, 15, 16
Very Good/High	≥ 1.35	5	1, 3, 7, 10

Regarding the item difficulty parameter (d), the values in this study range from −2.445 (Item 13) to +2.843 (Item 7). This distribution indicates that the instrument reflects a fairly wide range of difficulty levels, from very easy to very difficult. A good spread of item difficulty is essential to ensure that a test can measure students' abilities across different proficiency levels.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i2.90877

Most items are classified as very easy, as shown by d values below zero for more than 20 items, including Item 13 (d = −2.445), Item 24 (d = −6.647), and Item 26 (d = −3.152). These items are likely to be answered correctly even by examinees with low ability levels. Although such items may help reduce test anxiety and facilitate engagement at the beginning of the exam, an excessive number of very easy items reduces measurement precision. In particular, the ability to distinguish between medium- and high-ability examinees is limited, thereby weakening the instrument's overall discriminative strength.

Furthermore, several items belong to the easy category, such as Item 28 (d = 0.238) and Item 30 (d = 0.044). These items remain easy for most examinees; however, they begin to present some challenges for very low-ability students. Easy items, therefore, serve a purpose in differentiating between examinees at the lower end of the ability spectrum and those with average proficiency. Nevertheless, their quantity within this test appears limited, meaning their contribution to improving overall measurement balance is insufficient.

The instrument contains only a small number of medium-difficulty items, such as Item 31 (d = 0.494). Medium-difficulty items are considered ideal because they provide maximum information for examinees with average ability and contribute significantly to the test's general discriminative capacity. The limited number of items in this category indicates that the instrument does not yet optimally measure examinees across the full range of ability levels.

Difficult items are also present in the test, notably in Item 7 (d = 2.843) and Item 21 (d = 0.655). These items are typically answered correctly only by examinees with higher ability levels and are therefore essential for assessing more proficient students. However, extreme values, particularly the negative extreme value in Item 24 (d = −6.647), require careful review to ensure that they are not the result of technical issues, such as miskeyed responses or poor item construction.

Overall, the instrument tends to be too easy for the sampled examinee population, as evidenced by the dominance of low-difficulty items and the limited number of medium- and high-difficulty items. This imbalance negatively affects the test's ability to distinguish between examinees with moderate and high proficiency. Therefore, it is recommended to revise items with extreme difficulty values and to develop new items that fall within the medium-to-difficult range to strengthen the test's discriminative power and measurement quality (see Table 8).

Table 8. Item Difficulty Levels

Category	Range of Values	Example Items
Very Easy	$d < -2$	13, 24, 26
Easy	$-2 \leq d < 0$	28, 30
Moderate/Medium	$0 \leq d \leq 1$	31, 21
Difficult/Hard	$d > 1$	7, 10

Discussion

Analyzing the discrimination parameter (a) and item difficulty (d) within a test instrument is a critical step in determining its psychometric quality. IRT, particularly through the MIRT model implementation, offers a more comprehensive analytical lens because examinee ability is often composed of multiple dimensions rather than a single latent trait. Therefore, evaluating these parameters ensures that each item contributes appropriately to the overall measurement objective and supports the validity of score interpretations.

Item Sharpness in Differentiating Examinees (Discrimination Parameter)

Theoretically, the discrimination parameter indicates how well an item distinguishes between higher and lower ability examinees on a given dimension. In MIRT, discrimination is not expressed by a single parameter alone; it may be distributed across multiple ability

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i2.90877

dimensions (Embretson & Reise, 2013). High positive discrimination values indicate that an item is highly sensitive to changes in examinee ability. Conversely, negative discrimination values are regarded as indicative of problems with the item, whether in terms of substance, coding, or technical construction (Hambleton et al., 1991).

The findings show that most items have discrimination values in the −2.5 to +2.5 range, which is generally acceptable. However, 16 items have negative discrimination values. It is a serious concern since, conceptually, negative discrimination means that higher ability examinees are more likely to answer incorrectly, contrary to fundamental psychometric principles (Kasali & Adeyemi, 2022). Such a condition may indicate misalignment between item content and the intended ability construct, or technical issues such as miscoding of the answer key.

Several items, specifically Items 1, 3, 7, 10, and 22, exhibit high positive discrimination, even exceeding 1.35. Items with discrimination at this level are considered very good because they sharply differentiate examinees across ability levels (Zondo et al., 2021). Items of this kind form the backbone of a test instrument by providing maximal information around specific ability points.

In a multidimensional context that items showing discrimination on more than one dimension are not inherently problematic, as long as those dimensions are relevant to the construct being measured (Ayanwale et al., 2022). However, if an item shows negative discrimination on the principal dimension, as observed for Items 1 and 3, this may indicate errors in instrument design or coding (Jordan & Spiess, 2019). Thus, evaluating item quality cannot be separated from a substantive review of content and the role of each dimension in the theoretical construct.

The theoretical implications of negative discrimination are substantial. Such items can undermine construct validity by violating the monotonicity principle in IRT, stating that the probability of a correct response should increase as ability increases. This aligns with Karnia (2024), who stresses that test validity depends not only on internal reliability but also on the behavior of individual items. Moreover, items with low discrimination (e.g., < 0.35) should be revised or removed because they provide insufficient information for measurement (Sadeghi et al., 2025).

Variation and Distribution of Measured Ability (Item Difficulty)

The difficulty parameter (d) indicates the ability level at which an examinee has a 50% chance of answering an item correctly. In logistic models, d typically ranges from about −3 to +3. Values below −3 indicate very easy items, whereas values above +3 indicate very difficult items (Istiqlal et al., 2025). The analysis shows that d values in this instrument range from −6.647 to +2.843. This indicates a fairly wide spread of difficulty levels, but also reveals a dominance of very easy items. More than half of the items have $d < 0$, meaning they tend to be answered correctly by almost all examinees, including those at lower ability levels.

This condition carries important implications. While very easy items can help build examinees' confidence at the beginning of a test, an excessive number renders the instrument less informative for distinguishing among examinees at medium to higher ability levels. A shortage of medium to difficult items reduces the test's overall discriminative power. Previous research by Shafie et al. (2021) underscores the importance of a balanced distribution of difficulty so that the test information function spans the full ability spectrum. In other words, a well-designed test should comprise a mix of easy, medium, and difficult items to accurately measure low, average, and high-ability examinees. Perie (2020) also found that overly easy instruments tend to produce a ceiling effect, in which high-ability examinees cannot be effectively differentiated because too many achieve perfect or near-perfect scores.

Local research, such as Ernawati et al. (2024), indicates that in the context of educational assessment in Indonesia, a dominance of easy items tends to reduce the instrument's selectivity

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i2.90877

and diagnostic function. In the present instrument, the very limited number of medium and difficult items suggests that the measurement of examinees with medium to high ability is not yet optimal. This makes the instrument less suitable for academic selection or summative assessment purposes that require precise differentiation. Additionally, items with very extreme d values, either too low (< -3) or too high ($> +3$), should be re-examined because they tend to be statistically unstable and provide information only at a very narrow ability range (Hambleton et al., 1991). Such items generally contribute little to total test information and risk degrading measurement quality.

Based on the analysis above, several strategic steps are needed to improve instrument quality. First, items with negative discrimination and extreme difficulty should be examined qualitatively, including the answer key, wording, and content alignment with construct indicators. Item revision or deletion should be carried out systematically, considering empirical data and expert judgment (Diki & Yuliastuti, 2018). Second, new high-quality items need to be added. Ideally, such items should have discrimination above 0.65 and difficulty between 0.5 and 2.5 (Escudero et al., 2000). Items with these characteristics tend to provide maximal information for examinees with average to high ability, an ability range currently underrepresented in the instrument (Glas et al., 2003). Third, to strengthen instrument validity, further evaluation of model fit is required, including indices such as RMSEA, SRMSR, and the chi-square test. In addition, construct validity testing through confirmatory factor analysis (CFA) can provide further information on the alignment between the theoretical structure and empirical data. Differential Item Functioning (DIF) analysis is also recommended to ensure that there is no item bias across demographic groups, such as gender or educational background.

CONCLUSION

Analysis of discrimination and difficulty parameters in a test instrument based on Multidimensional Item Response Theory (MIRT) shows that item quality varies widely and requires careful attention to ensure valid and effective measurement. Regarding the discrimination parameter (a), most items fall within an acceptable range (-2.5 to $+2.5$); however, the presence of 16 items with negative discrimination indicates potential serious problems in content, coding, or item construction. Items with very high discrimination (e.g., > 1.35), such as Items 1, 3, 7, 10, and 22, are highly informative and constitute key components of the instrument. Negative discrimination, especially on the principal dimension, poses risks to construct validity and contradicts the basic IRT principle that the probability of a correct response should increase with ability.

With respect to item difficulty (d), the distribution is dominated by items categorized as easy ($d < 0$), while relatively few items fall into the medium and difficult categories. This may reduce the instrument's discriminative power, particularly for examinees with medium to high ability, and can lead to a ceiling effect. Extreme difficulty values (below -3 or above $+3$) also warrant caution because they contribute little to test information and may be statistically unstable.

The findings recommend that items with negative discrimination and extreme difficulty be reviewed qualitatively and then revised or removed. New, high-quality items (discrimination > 0.65 and difficulty between 0.5 and 2.5) should be added to better capture examinees with medium to high ability. Instrument validity should be reinforced through model fit testing (fit indices), construct validity (CFA), and checks for potential item bias (DIF analysis). Overall, these results underscore the importance of in-depth evaluation of each test item, both quantitatively and qualitatively, to ensure that the instrument provides valid, fair, and informative measurement across the spectrum of examinee ability.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i2.90877

Conflict of Interests

The authors declare that they have no conflict of interest related to this study.

REFERENCES

Ackerman, T. A., & Ma, Y. (2024). Examining differential item functioning from a multidimensional IRT perspective. *Psychometrika, 89*(1), 4-41. https://doi.org/10.1007/s11336-024-09965-6

Al-Qerem, W., Abdo, S., Jarab, A., Hammad, A., Eberhardt, J., Al-Asmari, F., . . . Zumot, R. (2025). Validation of the Arabic version of the Long-Term Conditions Questionnaire (LTCQ): A study of factor and Rasch analyses. *Healthcare, 13*(13), 1485. https://doi.org/10.3390/healthcare13131485

Alhamami, M. (2025). Intention over motivation: A holistic analysis of psychological constructs in Arabic as a foreign language learning. *Acta Psychologica, 258*, 105142. https://doi.org/10.1016/j.actpsy.2025.105142

Asadizanjani, N., Reddy Kottur, H., & Dalir, H. (2025). Testing and reliability in advanced packaging. In *Introduction to microelectronics advanced packaging assurance* (pp. 141-159). Springer. https://doi.org/10.1007/978-3-031-86102-4_8

Ayanwale, M. A., Chere-Masopha, J., & Morena, M. C. (2022). The classical test or item response measurement theory. *International Journal of Learning, Teaching and Educational Research, 21*(8), 384-406. https://doi.org/10.26803/ijlter.21.8.22

Belenguer, L. (2022). AI bias: Exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *AI and Ethics, 2*(4), 771-787. https://doi.org/10.1007/s43681-022-00138-8

Diki, D., & Yuliastuti, E. (2018). Discrepancy of difficulty level based on item analysis and test developers' judgment: Department of Biology at Universitas Terbuka, Indonesia. In D. Ifenthaler, A. Yuen, Y. An, & J. M. Spector (Eds.), *Educational technology to improve quality and access on a global scale: Papers from the Educational Technology World Conference (ETWC 2016), Indonesia* (pp. 215–225). Spriner Nature. https://doi.org/10.1007/978-3-319-66227-5_17

Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists.* Psychology Press. https://doi.org/10.4324/9781410605269

Engida, M. A., Iyasu, A. S., & Fentie, Y. M. (2024). Impact of teaching quality on student achievement: student evidence. *Frontiers in Education, 9*, 1367317. https://doi.org/10.3389/feduc.2024.1367317

Ernawati, E., Habibah, R. Y., Syarifah, N., Firmansyah, F., & Attamimi, H. a. R. (2024). Item analysis test of science, Indonesian language, and mathematics using the Rasch model in elementary schools. *Jurnal Penelitian dan Evaluasi Pendidikan, 28*(2), 195-209. https://doi.org/10.21831/pep.v28i2.75448

Escudero, E. B., Reyna, N. L., & Morales, M. R. (2000). The level of difficulty and discrimination power of the Basic Knowledge and Skills Examination (EXHCOBA). *Revista Electrónica de Investigación Educativa, 2*(1). http://redie.uabc.mx/vol2no1/contents-backhoff.html

Freed, R., McKinnon, D., Fitzgerald, M., & Norris, C. M. (2022). Development and validation of an astronomy self-efficacy instrument for understanding and doing. *Physical Review Physics Education Research, 18*(1), 010117. https://doi.org/10.1103/PhysRevPhysEducRes.18.010117

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i2.90877

Glas, C., Scheerens, J., & Thomas, S. M. (2003). *Educational evaluation, assessment, and monitoring: A systemic approach* (1st ed.). Taylor & Francis. https://doi.org/10.4324/9780203971055

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). SAGE.

Istiqlal, M., Putro, N. H. P. S., & Istiyono, E. (2025). Evaluating English language test items developed by teachers: An item response theory approach. *Voices of English Language Education Society, 9*(1), 218-230. https://doi.org/10.29408/veles.v9i1.27644

Jewsbury, P. A., & van Rijn, P. W. (2020). IRT and MIRT models for item parameter estimation with multidimensional multistage tests. *Journal of Educational and Behavioral Statistics, 45*(4), 383-402. https://doi.org/10.3102/1076998619881790

Jordan, P., & Spiess, M. (2019). Rethinking the interpretation of item discrimination and factor loadings. *Educational and Psychological Measurement, 79*(6), 1103-1132. https://doi.org/10.1177/0013164419843164

Jundi, M. (2023). Classical test theory in analyzing Arabic test questions: A descriptive study on item analysis research in Indonesia/الدراسة: نظرية الاختبار الكلاسيكية في تحليل الأسئلة العربية الوصفية على بحوث تحليل بنود الأسلة في إندونيسيا. *ATHLA: Journal of Arabic Teaching, Linguistic and Literature, 4*(2), 85-105. https://doi.org/10.22515/athla.v4i2.7747

Kadir, S., Sarif, S., & Fuadi, A. H. N. (2024). Item analysis of Arabic thematic questions to determine thinking level ability. *ELOQUENCE: Journal of Foreign Language, 3*(1), 28-41. https://doi.org/10.58194/eloquence.v3i1.1498

Karnia, R. (2024). Importance of reliability and validity in research. *Psychology and Behavioral Sciences, 13*(6), 137-141. https://doi.org/10.13140/RG.2.2.30985.45921

Kasali, J., & Adeyemi, A. A. (2022). Estimation of item parameter indices of NECO Mathematics multiple choice test items among Nigerian students. *Journal of Integrated Elementary Education, 2(*1), 43-54. https://doi.org/10.21580/jieed.v2i1.10187

Lee, W. C., Kim, S. Y., Choi, J., & Kang, Y. (2020). IRT approaches to modeling scores on mixed-format tests. *Journal of Educational Measurement, 57*(2), 230-254. https://doi.org/10.1111/jedm.12248

Madi, D., & Clinton, M. (2015). Rasch analysis of the Arabic language version of the functional disability inventory. *Journal of Pediatric Oncology Nursing, 32*(4), 230-239. https://doi.org/10.1177/1043454214554010

Mahmudi, I., Nurwardah, A., Rochma, S. N., & Nurcholis, A. (2023). Item analysis of Arabic language examination. *Ijaz Arabi Journal of Arabic Learning, 6*(3). https://doi.org/10.18860/ijazarabi.v6i3.19821

Mardapi, D. (2020). Assessing students' higher order thinking skills using multidimensional item response theory. *Problems of Education in the 21st Century, 78*(2), 196-214. https://doi.org/10.33225/pec/20.78.196

Ningsih, N. T. R., Rosidin, U., Viyanti, V., Distrik, I. W., & Abdurrahman, A. (2025). Development of an assessment instrument for students discipline and responsibility in physics practicum-based cooperative learning. *Sebatik, 29*(1), 67-73. https://doi.org/10.46984/sebatik.v29i1.2595

Ntumi, S. (2025). Advanced multidimensional item response theory modeling for high-stakes, cross-disciplinary competency assessments in sub-Saharan Africa: A psychometric

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i2.90877

approach to equity, adaptivity, and policy integration. *Research Square, Preprint*(Version 1). https://doi.org/10.21203/rs.3.rs-6418690/v1

Nury, A. H. A., Hikmah, H., & Masrun, M. (2025). Assessment Instruments for Tarkib and Mufrodat in the Ministry of Religion's Arabic language textbook. *Al-Lahjah: Jurnal Pendidikan, Bahasa Arab, dan Kajian Linguistik Arab, 8*(2), 1021-1031. https://doi.org/10.32764/lahjah.v8i2.5879

Oladele, J. I., & Ndlovu, M. (2021). A review of standardised assessment development procedure and algorithms for computer adaptive testing: Applications and relevance for fourth industrial revolution. *International Journal of Learning, Teaching and Educational Research, 20*(5), 1-17. https://www.ijlter.org/index.php/ijlter/article/view/3551

Pardede, T., Santoso, A., Diki, D., Retnawati, H., Rafi, I., Apino, E., & Rosyada, M. N. (2023). Gaining a deeper understanding of the meaning of the carelessness parameter in the 4PL IRT model and strategies for estimating it. *REID (Research and Evaluation in Education)*, *9*(1), 86–117. https://doi.org/10.21831/reid.v9i1.63230

Perie, M. (2020). Comparability across different assessment systems. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations,* pp. 123-148. National Academy of Education. https://naeducation.org/wp-content/uploads/2020/06/Comparability-of-Large-Scale-Educational-Assessments.pdf

Puia, A.-M., Mihalcea, A., & Rotărescu, V. Ş. (2025). Well-being factors. An item-level analysis of the positive cognitive triad role, in the relationship between resilience and well-being. *Acta Psychologica, 253*, 104692. https://doi.org/10.1016/j.actpsy.2025.104692

Ramadhan, M. R., & Subando, J. (2025). Analisis kualitas butir soal fiqih dan kemampuan siswa di Madrasah Aliyah Negeri 1 Surakarta. *Al Ulum Jurnal Pendidikan Islam*, *5*(2), 126-138. https://doi.org/10.54090/alulum.698

Sadeghi, P., Pourabbas, A., Dehghani, G., & Katebi, K. (2025). Quantitative and qualitative item analysis of exams of basic medical sciences departments of Tabriz University of Medical Sciences in 2023. *BMC Medical Education, 25*(1), 937. https://doi.org/10.1186/s12909-025-07539-3

Saepudin, S., Pabbajah, M. T. H., & Pabbajah, M. (2024). Unleashing the power of reading: Effective strategies for non-native Arabic language learners. *Alsinatuna, 9*(2), 109-130. https://doi.org/10.28918/alsinatuna.v9i2.7826

Shafie, S., Majid, F. A., Hoon, T. S., & Damio, S. M. (2021). Evaluating construct validity and reliability of intention to transfer training conduct instrument using Rasch model analysis. *Pertanika Journal of Social Sciences & Humanities, 29*(2), 1055–1070. https://doi.org/10.47836/pjssh.29.2.17

Stalikas, A., Triliva, S., & Roussi, P. (2018). Exploratory factor analysis. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of personality and individual differences*. Springer. https://doi.org/10.1007/978-3-319-28099-8_1385-1

Terry, D., & Nguyen, H. (2024). Assessing measuring instruments. In D. Whitehead & D. Terry (Eds.), *Nursing and midwifery research: Methods and appraisal for evidence based practice* (7th ed.), pp. 151-167. Elsevier. https://www.elsevierhealth.com.au/nursing-and-midwifery-research-9780729544665.html

Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology, 44*(3), 219–246. https://doi.org/10.1177/0095798418771807

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i2.90877

Wilson, M. (2023). *Constructing measures: An item response modeling approach* (2nd ed.). Routledge. https://doi.org/10.4324/9781003286929

Zakkiyah, M. Y., Fidyahwati, N. M., Ma'suq, A. T., & Anggraini, N. (2024). Assessment design and analysis of Arabic reading skills instructional materials. *IJIE International Journal of Islamic Education, 3*(1), 31-46. https://doi.org/10.35719/ijie.v3i1.2000

Zeinoun, P., Iliescu, D., & El Hakim, R. (2022). Psychological tests in Arabic: A review of methodological practices and recommendations for future use. *Neuropsychology Review, 32*(1), 1-19. https://doi.org/10.1007/s11065-021-09476-6

Zondo, N. P., Zewotir, T., & North, D. E. (2021). The level of difficulty and discrimination power of the items of the National Senior Certificate Mathematics Examination. *South African Journal of Education, 41*(4), 1-13. https://doi.org/10.15700/saje.v41n4a1935

ISSN (print) 2685-7111 | ISSN (online) 2338-6061