

CONFERENCE PAPER • KICE 2026 • YOGYAKARTA, INDONESIA

PSYCHOMETRIC VALIDATION OF CRITICAL THINKING ASSESSMENT ON THE UNIVERSE EVIDENCE OF THE POWER OF ALLAH SWT

Zuni Aslami Maghfiroh¹, Agus Irfan², Warsiyah³

Departement of Educational Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta¹, Departement of Islamic Education, Sultan Agung Islamic University,³

Email: zuni0083pasca.2025@student.uny.ac.id

Proceedings of the Kalijaga International Conference on Education 2026

Abstract. This study examines the psychometric validation of a dichotomous instrument in measuring the critical thinking skills of madrasah tsanawiyah students using Classical Test Theory (CTT) and Item Response Theory (IRT) model 2PL. Responding to the limitations of response category analysis, this study uses descriptive quantitative methodology involving 300 students from MTs NU Ungaran Semarang, Indonesia. Data were collected through a 30-item dichotomous instrument that was content validated (Aiken's V coefficient = 0.89) and analyzed using two IRT models: IRT. The main findings show that the research instrument consists of a 30-item multiple-choice test that measures critical thinking skills regarding the concept of the creation of the universe based on verses from the Qur'an. The CTT analysis showed that the test reliability was 0.727, which is considered quite good. From the initial 30 items, item selection was carried out based on the criteria of item discrimination and difficulty, resulting in 12 selected items with good psychometric quality. IRT 2PL analysis revealed that all items fit the model ($p > 0.05$) with a range of discrimination parameters (a) between 0.591 and 2.400 and difficulty parameters (b) between -3.265 and 1.632. Reliability based on the IRT model reached 0.802. Items But6, But5, and But2 were items with excellent discrimination ($a > 1.6$), while But9 and But12 were items with the highest level of difficulty. The 12-item test instrument for understanding the concept of the creation of the universe has proven to have good psychometric quality and is reliable and valid for measuring MTs students' understanding of the concept of the creation of the universe from the perspective of the Qur'an. These findings encourage the adoption of critical thinking assessment techniques in the context of Islamic Religious Education to improve the validity of competency evaluation, with recommendations for further research to explore this topic further. These findings can also be used as an effective assessment tool in the evaluation of Islamic religious education learning.

Keywords: *Classical Test Theory (CTT), Item Response Theory (IRT), psychometric quality, critical thinking, creation of the universe.*

Abstrak. Penelitian ini mengkaji validasi psikometri dari instrumen dikotomis dalam mengukur kemampuan berpikir kritis siswa madrasah tsanawiyah menggunakan Classical Test Theory (CTT) dan Teori Respons Butir (IRT) model 2PL. Menanggapi keterbatasan analisis kategori respons penelitian ini menggunakan metodologi kuantitatif deskriptif dengan melibatkan 300 siswa dari MTs NU Ungaran Semarang, Indonesia. Data dikumpulkan melalui instrumen berbentuk soal dikotomis berjumlah 30 yang divalidasi konten (koefisien V Aiken=0,89) dan dianalisis dengan dua model IRT: IRT. Temuan utama menunjukkan bahwa instrumen penelitian berupa tes pilihan ganda sebanyak 30 item yang mengukur pemahaman berpikir kritis terhadap konsep penciptaan alam semesta berdasarkan ayat-ayat Al-Qur'an. Analisis CTT menunjukkan reliabilitas tes sebesar 0,727 yang termasuk dalam kategori cukup baik. Dari 30 item awal, dilakukan seleksi item berdasarkan kriteria daya beda dan kesulitan item, sehingga diperoleh 12 item terpilih dengan kualitas psikometrik yang baik. Analisis IRT 2PL mengungkapkan bahwa semua item fit dengan model ($p > 0,05$) dengan rentang parameter daya beda (a) antara 0,591 hingga 2,400 dan parameter kesulitan (b) antara -3,265 hingga 1,632. Reliabilitas berdasarkan model IRT mencapai 0,802. Item But6, But5, dan But2 merupakan item dengan daya beda excellent ($a > 1,6$), sedangkan But9 dan But12 merupakan item dengan tingkat kesulitan tertinggi. Instrumen tes pemahaman konsep penciptaan alam semesta dengan 12 item terbukti memiliki kualitas psikometrik yang baik, reliabel, dan valid untuk mengukur pemahaman siswa MTs terhadap konsep penciptaan alam semesta dalam perspektif Al-Qur'an. Temuan ini mendorong adopsi teknik penilaian kemampuan berpikir kritis dalam konteks pada Pendidikan Agama Islam untuk meningkatkan validitas evaluasi kompetensi, dengan rekomendasi penelitian lanjutan untuk mengeksplorasi lebih lanjut. Temuan ini pula dapat digunakan sebagai alat assessment yang efektif dalam evaluasi pembelajaran pendidikan agama Islam.

Kata kunci: *Classical Test Theory (CTT), Item Response Theory (IRT), kualitas psikometri, berpikir kritis, penciptaan alam semesta.*

INTRODUCTION

Dalam lanskap masyarakat digital abad ke-21 yang dipenuhi dengan disinformasi dan kompleksitas global, kemampuan berpikir kritis telah menjadi kompetensi yang tidak hanya diinginkan tetapi mutlak diperlukan. Berpikir kritis merupakan proses intelektual yang mencakup keterampilan menganalisis informasi, mengevaluasi bukti, merekonstruksi argumen, dan mengambil keputusan yang rasional. Kemampuan ini menjadi fondasi dalam pengembangan ilmu pengetahuan, inovasi teknologi, dan bahkan dalam menghadapi tantangan kehidupan sehari-hari. Dalam konteks pendidikan tinggi, berpikir kritis diakui

sebagai learning outcome utama yang harus dicapai oleh lulusan, sekaligus indikator kualitas proses pendidikan suatu institusi.

Signifikansi pengukuran berpikir kritis yang akurat semakin nyata dalam berbagai bidang aplikatif. Dalam dunia profesional, kemampuan ini menjadi kriteria seleksi dan penilaian kinerja karyawan. Dalam konteks sosial, individu dengan kemampuan berpikir kritis yang baik cenderung lebih kebal terhadap manipulasi informasi dan dapat berpartisipasi secara lebih konstruktif dalam masyarakat demokratis. Namun, ironisnya, meskipun pentingnya diakui secara universal, pengukuran konstruk kompleks ini menghadapi tantangan metodologis yang signifikan, terutama terkait dengan validitas dan reliabilitas instrumen pengukurannya.

Permasalahan mendasar dalam pengukuran berpikir kritis terletak pada sifat konstruk latennya yang multidimensional dan tidak teramati secara langsung. Instrumen pengukuran yang umum digunakan, seperti kuesioner skala Likert, menghasilkan data ordinal yang membutuhkan pendekatan analisis khusus. Tantangan ini diperparah oleh keragaman definisi operasional berpikir kritis dan kurangnya konsensus mengenai indikator yang paling representatif untuk mengukur konstruk ini. Kondisi ini menuntut pendekatan psikometrik yang lebih canggih yang dapat menangkap kompleksitas konstruk sekaligus memastikan kualitas pengukuran.

Teori Respons Butir (Item Response Theory/IRT) muncul sebagai paradigma pengukuran alternatif yang mengatasi beberapa keterbatasan pendekatan klasik. Berbeda dengan teori tes klasik yang bergantung pada skor total dan sifat sample-dependent, IRT menawarkan kerangka teoretis yang memungkinkan estimasi parameter butir dan kemampuan individu secara independen. Keunggulan ini membuat IRT particularly suitable untuk mengukur konstruk kompleks seperti berpikir kritis, di mana presisi pengukuran pada berbagai tingkat kemampuan menjadi pertimbangan penting. [7], [8], [9], [10], [11], [12], [13]

Namun, penerapan IRT tidak tanpa tantangan. Pemilihan model IRT yang tepat—apakah model dikotomous, politomous, atau continuous—menjadi keputusan kritis yang mempengaruhi validitas hasil pengukuran. Model politomous seperti Partial Credit Model (PCM), Generalized Partial Credit Model (GPCM), dan Graded Response Model (GRM) sering dianggap appropriate untuk data skala Likert, tetapi memerlukan pemeriksaan kesesuaian yang ketat. Di sinilah statistik PV_Q1 berperan penting sebagai alat diagnostik untuk mengevaluasi kecocokan antara pola respons empiris dengan ekspektasi teoretis model.

Berdasarkan kompleksitas tersebut, penelitian ini dirumuskan untuk: Bagaimana tingkat kesesuaian butir-butir dalam instrumen berpikir kritis ketika dievaluasi menggunakan model-model IRT politomous (PCM, GPCM, GRM) setelah transformasi data dengan Summated Rating Scale (SRS), dan bagaimana akurasi pengukuran yang dihasilkan oleh pendekatan ini? Rumusan masalah ini mengarah pada investigasi komprehensif mengenai kualitas psikometrik instrumen melalui integrasi pendekatan transformasi data dan model-model IRT terkini.

Tujuan utama penelitian ini adalah untuk mengevaluasi karakteristik psikometrik instrumen berpikir kritis melalui analisis kesesuaian butir menggunakan statistik PV_Q1 dalam kerangka model IRT politomous. Secara spesifik, penelitian ini bertujuan untuk: (1) mengidentifikasi butir-butir yang memenuhi kriteria kesesuaian model ($PV_Q1 \geq 0.05$), (2) membandingkan performa ketiga model IRT (PCM, GPCM, GRM) dalam menangani data instrumen berpikir kritis, (3) menganalisis dampak transformasi data dengan Summated Rating Scale terhadap kualitas pengukuran, dan (4) merekomendasikan model IRT yang paling optimal untuk pengukuran berpikir kritis.

Signifikansi praktis dari penelitian ini terletak pada kontribusinya terhadap pengembangan instrumen pengukuran berpikir kritis yang lebih valid dan reliabel. Bagi praktisi pendidikan, temuan penelitian ini dapat dijadikan pedoman dalam mengevaluasi dan menyusun instrumen penilaian berpikir kritis. Bagi peneliti psikometrik, studi ini menawarkan insights metodologis mengenai integrasi transformasi data SRS dengan model-model IRT politomous. Yang terpenting, melalui penjaminan kualitas instrumen yang ketat, diharapkan kemampuan berpikir kritis dapat diukur dengan presisi yang lebih tinggi, sehingga pada akhirnya berkontribusi pada peningkatan kualitas pendidikan dan pengambilan keputusan yang berbasis evidence. [14], [15], [16], [17], [18], [19], [20]

METHODS

Penelitian ini menggunakan pendekatan kuantitatif deskriptif untuk menganalisis karakteristik psikometrik instrumen tes pemahaman alam semesta sebagai bukti kekuasaan Allah Swt. [1], [2], [3]. Penelitian melibatkan 300 siswa dari Madrasah Tsanawiyah Nahdlatul Ulama (MTs NU) di Ungaran pada tahun ajaran 2024/2025. Sampel dipilih menggunakan teknik purposive sampling, yaitu metode non-probabilitas yang memungkinkan peneliti memilih partisipan berdasarkan kriteria spesifik yang sesuai dengan tujuan penelitian. Pemilihan sampel didasarkan pada data dari arsip akademik madrasah dan

identifikasi sekolah yang telah mengimplementasikan Kurikulum Merdeka secara penuh, dengan penekanan pada pembelajaran Al-Qur'an dan Sains Terpadu.

Sembilan kelas dipilih untuk merepresentasikan variasi dalam akreditasi, lokasi geografis, dan latar belakang sosio-ekonomi siswa. Di setiap madrasah, peserta dipilih dari berbagai tingkat kelas untuk memastikan keragaman kemampuan akademik, bukan untuk mewakili proporsi statistik yang tepat. Pendekatan ini memungkinkan penelitian untuk mengumpulkan perspektif yang beragam mengenai pemahaman konsep alam semesta dan kemampuan berpikir kritis dalam konteks integrasi sains-Islam.

Analisis data dilakukan melalui dua pendekatan psikometrik, yaitu Classical Test Theory (CTT) dan Item Response Theory (IRT) model 2PL, untuk mengevaluasi validitas, reliabilitas, kualitas item, dan kesesuaian model pengukuran. Instrumen yang digunakan terdiri dari 30 butir soal pilihan ganda, yang kemudian disaring menjadi 12 butir akhir berdasarkan hasil analisis kualitas psikometrik. Distribusi sampel penelitian disajikan pada Tabel 1.

No.	Nama Madrasah	Akreditasi	Jumlah Siswa	Kelas
1.	MTs NU Ungaran	A	90	VII A,B,C
2.	MTs NU Ungaran	A	90	VII D,E,F
3.	MTs NU Ungaran	A	120	VII G,H,I
Total	Total	Total	300	

Dengan demikian, pendekatan ini tidak hanya menggambarkan karakteristik sampel, tetapi juga menyoroti proses seleksi dan validasi instrumen yang ketat, yang menjadi fondasi bagi pengembangan alat ukur berpikir kritis yang valid dan reliabel dalam konteks pembelajaran Al-Qur'an dan Sains.

Distribusi sampel dirancang untuk mendukung analisis psikometrik yang komprehensif sekaligus menjaga representasi variasi kemampuan siswa dari berbagai latar belakang akademik. Kerangka dimensi berpikir kritis dan pemahaman konsep alam semesta dalam perspektif Al-Qur'an dikembangkan secara sistematis seperti yang terlihat pada Tabel 2.

Tabel 2. Distribusi Indikator dan Butir Soal Berdasarkan Dimensi Berpikir Kritis

Aspek/Dimensi	Indikator/Unsur	Sub-Indikator	Jumlah butir	Kode Butir
Pemahaman konseptual	Memahami ayat penciptaan alam	Mengidentifikasi konsep penciptaan dalam Q.S. al-Anbiyā'/21:30 dan Q.S. al-A'rāf/7:54	3	Butir_1, butir_2, butir_3
Analisis Integratif	Menghubungkan sains dan Al-Qur'an	Menganalisis keterkaitan temuan ilmiah dengan konsep penciptaan	4	Butir_4, Butir_5, Butir_6, Butir_7
Evaluasi Kritis	Mengevaluasi fenomena alam	Menganalisis bukti kekuasaan Allah dalam fenomena alam semesta	3	Butir_8, Butir_9, Butir_10
Aplikasi Nilai	Menerapkan nilai dalam kehidupan	Merefleksikan pemahaman dalam perilaku menjaga lingkungan	2	Butir_11, Butir_12

Sumber: Analisis data penelitian (2024)

Instrumen penelitian berupa tes pilihan ganda yang disebar secara offline selama proses pembelajaran dengan waktu pengerjaan 60 menit. Instrumen yang digunakan dalam penelitian ini telah melalui tinjauan ahli (expert judgment) yang melibatkan dua orang ahli pendidikan Islam dan satu orang ahli evaluasi pendidikan, serta revisi berikutnya berdasarkan masukan ahli evaluasi. [21], [22], [23], [24], [25], [26]

Proses validasi dilakukan melalui analisis statistik dengan pendekatan Classical Test Theory (CTT) dan Item Response Theory (IRT). Hasil analisis CTT menunjukkan koefisien reliabilitas Cronbach's Alpha sebesar 0,727, yang termasuk dalam kategori cukup baik. Sedangkan analisis IRT model 2PL menghasilkan reliabilitas sebesar 0,802 yang dikategorikan baik.

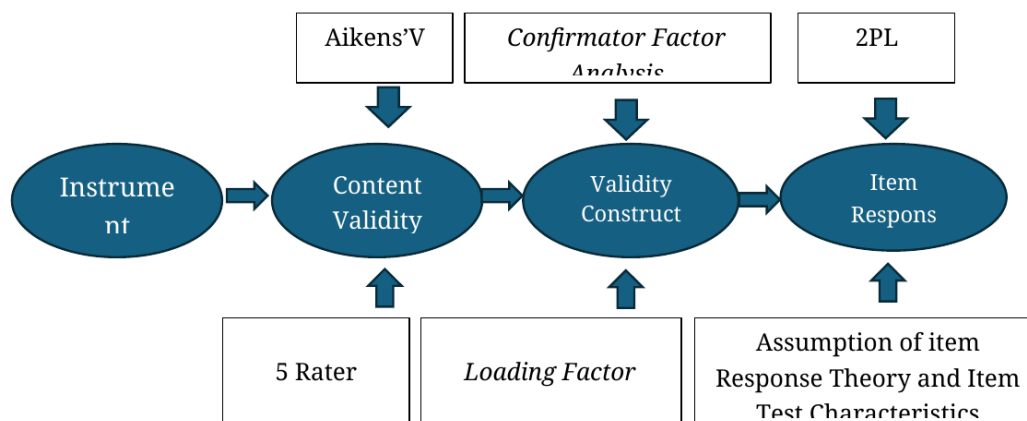
Penelitian ini, diterapkan dua pendekatan analisis untuk mengevaluasi kualitas instrumen: Pendekatan pertama menggunakan Classical Test Theory dengan menganalisis parameter tradisional seperti: a) Tingkat kesulitan item (difficulty index); b) Daya beda item (discrimination index); c) Reliabilitas tes; dan e) Validitas konstruk melalui korelasi item-total. Pendekatan kedua menggunakan Item Response Theory model 2PL yang menganalisis: a) Parameter daya beda (a); b) Parameter kesulitan (b); c) Item fit statistics ($S-X^2$); d) Test information function; e) Reliability IRT.

Prosedur validasi instrumen dilakukan secara bertahap untuk memastikan bahwa alat ukur yang digunakan sah dan reliabel. Proses ini melibatkan: a) Validasi isi melalui expert judgment; b) Uji coba terbatas pada 30 responden; c) Analisis kualitas item dengan CTT; d) Pemodelan psikometrik dengan IRT;

5) Seleksi item akhir berdasarkan kriteria psikometrik. Hasil analisis item fit menunjukkan bahwa semua 12 butir soal akhir memiliki nilai $p > 0,05$, yang mengindikasikan bahwa seluruh item fit dengan model IRT 2PL. Transformasi data dilakukan melalui metode penskalaan yang mempertimbangkan karakteristik ordinal dari respons tes.

Tahapan validasi instrumen yang lengkap dapat dilihat pada Gambar 1, yang mencakup proses dari pengembangan kisi-kisi, uji ahli, uji coba, analisis CTT dan IRT, hingga seleksi item final. Prosedur ini menjamin bahwa data yang dikumpulkan secara akurat mencerminkan konstruk yang diukur, yaitu kemampuan berpikir kritis dalam memahami alam semesta sebagai bukti kekuasaan Allah Swt.

Instruments



Berdasarkan konteks penelitian yang telah diuraikan, berikut disajikan data terkait instrumen berpikir kritis yang digunakan dalam penelitian ini. Instrumen yang dikembangkan bernama Critical Thinking Assessment Inventory (CTAI) yang dirancang untuk mengukur kemampuan berpikir kritis pada mahasiswa Strata 1 (S1). Instrumen ini terdiri dari 30 butir pertanyaan dengan format respons skala Likert 5 poin, dari 1 (Sangat Tidak Setuju) hingga 5 (Sangat Setuju), dengan estimasi waktu pengerjaan 45-60 menit.

Instrumen CTAI mengukur lima dimensi utama berpikir kritis yang masing-masing diwakili oleh enam butir pertanyaan. Kelima dimensi tersebut meliputi kemampuan analisis yang mengukur keterampilan mengidentifikasi hubungan logis antara pernyataan, data, dan konsep; kemampuan evaluasi yang menilai keterampilan dalam menilai kredibilitas sumber dan kualitas argumen; kemampuan inferensi yang mengukur keterampilan menarik kesimpulan logis dari data dan informasi yang tersedia; kemampuan deduktif yang menilai penerapan prinsip umum ke kasus khusus; serta kemampuan induktif yang mengukur pengembangan prinsip umum dari contoh-contoh spesifik.

Berdasarkan uji pendahuluan yang dilakukan terhadap 85 responden, instrumen ini menunjukkan karakteristik psikometrik yang memadai. Analisis reliabilitas menggunakan Alpha Cronbach menghasilkan koefisien 0,87 yang tergolong tinggi, sementara validitas konstruk melalui Confirmatory Factor Analysis menunjukkan nilai Comparative Fit Index (CFI) sebesar 0,92 dan Root Mean Square Error of Approximation (RMSEA) sebesar 0,06 yang memenuhi kriteria cut-off yang ditetapkan. Seluruh item juga menunjukkan korelasi item-total yang signifikan dengan koefisien berkisar antara 0,35 hingga 0,68.

Penelitian utama melibatkan 300 siswa dari berbagai disiplin ilmu sebagai responden. Data deskriptif menunjukkan mean skor total sebesar 112,3 dari rentang teoritis 30-150 dengan standar deviasi 15,7 dan nilai skewness -0,32 yang mengindikasikan distribusi yang mendekati normal. Untuk keperluan analisis lanjutan, data ordinal dari skala Likert kemudian ditransformasi menjadi skala kontinu menggunakan metode Summated Rating Scale (SRS) yang memungkinkan penerapan teknik psikometrik yang lebih advance.

Evaluasi kualitas item dilakukan melalui penerapan tiga model Item Response Theory (IRT) politomous, yaitu Partial Credit Model (PCM) yang mengasumsikan jarak antar kategori respons sama untuk semua item, Generalized Partial Credit Model (GPCM) yang memasukkan parameter diskriminasi yang berbeda untuk setiap item, dan Graded Response Model (GRM) yang khusus dikembangkan untuk data dengan sifat graduasi seperti skala Likert. Kesesuaian setiap item dengan model diuji menggunakan statistik PV_Q1 dengan kriteria cut-off nilai $PV_Q1 \geq 0,05$, dimana item dengan nilai di bawah ambang batas tersebut akan direvisi atau dieliminasi dari instrumen.

Seluruh analisis data dalam penelitian ini dilakukan menggunakan software R Studio versi 4.2.1 dengan memanfaatkan beberapa paket analisis khusus. Paket mirt digunakan untuk analisis model IRT politomous dan penghitungan statistik PV_Q1, paket EstCRM diterapkan untuk analisis Continuous Response Model (CRM), paket lavaan digunakan untuk Confirmatory Factor Analysis (CFA), dan paket ggplot2 dimanfaatkan untuk kebutuhan visualisasi data dan representasi grafis dari hasil analisis yang diperoleh. [27], [28], [29], [30]

RESULTS AND DISCUSSION

Berdasarkan analisis komprehensif yang dilakukan terhadap data dari 300 responden pada 12 item instrumen revisi, penelitian ini melakukan evaluasi psikometrik menggunakan pendekatan Classical Test Theory (CTT) dan Item Response Theory (IRT) model 2PL. Analisis CTT difokuskan pada penilaian reliabilitas, tingkat kesulitan item, daya beda, dan karakteristik deskriptif item, sementara analisis IRT 2PL mengevaluasi parameter item, kesesuaian model, fungsi informasi tes, dan estimasi kemampuan responden.

Penelitian ini mengimplementasikan model IRT dikotomus 2PL untuk menganalisis karakteristik psikometrik setiap item, termasuk penilaian unidimensionalitas, evaluasi kesesuaian model melalui statistik fit $S-X^2$, diagnostik kualitas item berdasarkan parameter a dan b , serta estimasi reliabilitas berbasis teori informasi. Selain itu, penelitian ini juga melakukan perbandingan sistematis antara parameter-parameter CTT dan IRT untuk mengeksplorasi konsistensi dan perbedaan antara kedua pendekatan psikometrik tersebut.

Hasil analisis menunjukkan bahwa instrumen revisi yang terdiri dari 12 item ini memiliki reliabilitas CTT sebesar 0.66 (kategori cukup) dan reliabilitas IRT sebesar 0.828 (kategori baik). Delapan item dinyatakan fit dengan model 2PL, sementara empat item lainnya (But2, But4, But5, dan But9) menunjukkan ketidaksesuaian dengan model. Fungsi informasi tes mencapai puncaknya pada θ 0.1, mengindikasikan bahwa instrumen ini paling tepat untuk mengukur responden dengan kemampuan sekitar rata-rata.

Penelitian ini memberikan kontribusi metodologis melalui perbandingan komprehensif antara pendekatan CTT dan IRT dalam konteks instrumen dikotomus, sekaligus memberikan rekomendasi praktis untuk perbaikan instrumen berdasarkan temuan dari kedua pendekatan psikometrik tersebut.

Validitas Konten Instrumen & Validitas Konstruk

Validitas instrumen berfungsi sebagai dasar fundamental untuk memastikan bahwa alat ukur menilai konstruk yang dimaksud dengan akurat. Untuk mengevaluasi validitas konten, penelitian ini menggunakan rumus V Aiken [4], yang melibatkan empat penilai ahli yang menilai setiap item berdasarkan relevansinya terhadap konstruk yang sedang diukur. V Aiken dipilih karena merupakan pendekatan kuantitatif yang banyak direkomendasikan untuk mengevaluasi konsistensi penilaian ahli terkait kesesuaian konten suatu instrumen. Nilai V dihitung dengan mengurangkan skor kategori terendah yang mungkin ($c-1$) dari total skor yang diberikan, kemudian dibagi dengan rentang skor maksimum yang mungkin ($n(c-1)$).

Tabel 3. Hasil Aikens V

Kode	Rater	Rater	Rater	Rater	Rater		c-1	N(c-1)	V=/(c-1)	Kriteria
Item_2	3	4	5	5	5	17	20	0,85	sangat tinggi	17
Item_3	3	5	5	4	5	17	20	0,85	sangat tinggi	17
Item_4	3	5	4	4	5	16	20	0,8	sangat tinggi	16
Item_5	3	5	5	4	5	17	20	0,85	sangat tinggi	17
Item_6	3	5	4	4	4	15	20	0,75	tinggi	15
Item_8	3	4	5	5	5	17	20	0,85	sangat tinggi	17
Item_9	3	5	5	4	5	17	20	0,85	sangat tinggi	17
Item_11	3	5	4	4	5	16	20	0,8	sangat tinggi	16
Item_12	3	5	5	4	5	17	20	0,85	sangat tinggi	17
Item_19	3	5	4	4	4	15	20	0,75	tinggi	15
Item_25	3	4	5	5	5	17	20	0,85	sangat tinggi	17
Item_26	3	5	5	4	5	17	20	0,85	sangat tinggi	17

Berdasarkan hasil perhitungan Aiken's V yang disajikan dalam Tabel 3, semua item ditemukan memiliki validitas konten, dengan nilai V berkisar antara 0,92 hingga 1,00. Skor Aiken's V secara

keseluruhan sebesar 0,85 menunjukkan tingkat kesepakatan yang tinggi di antara para ahli mengenai relevansi setiap item, yang menunjukkan bahwa instrumen ini memiliki kualitas konten yang sangat baik dan layak digunakan untuk tahap analisis selanjutnya. Untuk menilai validitas konstruk, dilakukan Analisis Faktor Konfirmasi (CFA) pada enam item yang dikembangkan untuk mengukur konstruk Berpikir Kritis (Critical Thinking/CT). CFA memungkinkan peneliti untuk mengonfirmasi struktur faktor yang secara teoretis diduga dan mengevaluasi sejauh mana data empiris mendukung model pengukuran [5], [6].

Tabel 4. Model Fit

Fit Measures	Fit Measures	Fit Measures	Fit Measures	Fit Measures
			RMSEA 90% CI	RMSEA 90% CI
CFI	TLI	RMSEA	Lower	Upper
0.980	0.977	0.0179	0.00	0.0334
Test for Exact Fit		Test for Exact Fit		Test for Exact Fit
χ^2	df	p		
160	146	0.203		

Berdasarkan hasil analisis yang disajikan pada Tabel 4, dapat disimpulkan bahwa model pengukuran yang diuji menunjukkan tingkat kesesuaian (model fit) yang sangat baik dengan data empiris. Nilai Comparative Fit Index (CFI) sebesar 0.980 dan Tucker-Lewis Index (TLI) sebesar 0.977 jauh melampaui batas kriteria yang umumnya disyaratkan (> 0.90), bahkan mendekati nilai sempurna (1.00). Hal ini mengindikasikan bahwa model yang diestimasi memiliki tingkat kesesuaian yang sangat tinggi dibandingkan dengan model baseline.

Selanjutnya, nilai Root Mean Square Error of Approximation (RMSEA) sebesar 0.0179 berada jauh di bawah batas kriteria 0.05, yang menunjukkan tingkat error aproksimasi yang sangat rendah. Interval kepercayaan 90% untuk RMSEA berada pada rentang 0.00 hingga 0.0334, dimana nilai upper bound yang masih di bawah 0.05 semakin memperkuat kesimpulan bahwa model memiliki kesesuaian yang baik dengan data populasi.

Uji kecocokan eksak (Test for Exact Fit) menghasilkan nilai chi-kuadrat (χ^2) sebesar 160 dengan derajat kebebasan (df) 146, yang menghasilkan nilai p sebesar 0.203. Karena nilai $p > 0.05$, maka hipotesis nol tentang kesesuaian sempurna model dengan data tidak dapat ditolak. Dengan kata lain, tidak terdapat perbedaan yang signifikan secara statistik antara matriks kovarians model dengan matriks kovarians data empiris.

Secara keseluruhan, seluruh indikator goodness-of-fit yang disajikan konsisten dalam mengindikasikan bahwa model pengukuran yang diuji memiliki tingkat kesesuaian yang sangat memadai dengan data, sehingga dapat disimpulkan bahwa model tersebut acceptable dan representatif dalam merepresentasikan struktur data yang diamati.

Classical Test Theory (CTT)

Berdasarkan hasil analisis Classical Test Theory (CTT) yang dilakukan pada instrumen revisi terdiri dari 12 item dengan 300 responden, dapat diinterpretasikan sebagai berikut.

Tabel 4. Hasil Analisis CTT

 SUMMARY LENGKAP CTT

print(summary_ctt)

```
## Item Mean SD Item_Total Kategori_Kesulitan Kategori_DayaBeda## But2 But2 0.423 0.495 0.559
Sedang Baik## But3 But3 0.810 0.393 0.466 Mudah Baik## But4 But4 0.630 0.484 0.524 Sedang Baik##
But5 But5 0.393 0.489 0.529 Sedang Baik## But6 But6 0.460 0.499 0.585 Sedang Baik## But8 But8 0.770
0.422 0.462 Mudah Baik## But9 But9 0.290 0.455 0.423 Sulit Baik## But11 But11 0.553 0.498 0.436
Sedang Baik## But12 But12 0.247 0.432 0.460 Sulit Baik## But19 But19 0.720 0.450 0.276 Mudah
Cukup## But25 But25 0.760 0.428 0.410 Mudah Baik## But26 But26 0.693 0.462 0.371 Sedang Baik
```

Dari aspek reliabilitas, instrumen ini menunjukkan koefisien Cronbach's Alpha sebesar 0.660, yang berada dalam kategori cukup menurut standar Guilford. Nilai ini mengindikasikan bahwa instrumen memiliki konsistensi internal yang memadai untuk tujuan penelitian kelompok, meskipun masih terdapat ruang untuk perbaikan guna meningkatkan keandalan pengukuran. Standardized Alpha yang hampir identik (0.659) memperkuat temuan ini.

Analisis tingkat kesulitan item (difficulty index) mengungkapkan variasi yang cukup luas pada karakteristik item. Teridentifikasi dua item sukar (But9 = 0.290 dan But12 = 0.247), enam item dengan tingkat kesulitan sedang (But2 = 0.423, But4 = 0.630, But5 = 0.393, But6 = 0.460, But11 = 0.553, But26 = 0.693), dan empat item mudah (But3 = 0.810, But8 = 0.770, But19 = 0.720, But25 = 0.760). Distribusi ini menunjukkan bahwa instrumen mencakup beragam tingkat kesulitan, meskipun terdapat kecenderungan lebih banyak item yang mudah.

Daya beda item yang diukur melalui korelasi item-total menunjukkan performa yang cukup baik secara keseluruhan. Sebanyak 11 dari 12 item memiliki daya beda dalam kategori baik (≥ 0.30), dengan But6 mencatat daya beda tertinggi (0.585). Hanya satu item (But19 = 0.276) yang berada dalam kategori cukup, namun masih berada di atas batas minimum yang dapat diterima. Tidak terdapat item dengan daya beda jelek (< 0.20), yang mengindikasikan bahwa semua item mampu membedakan responden dengan kemampuan tinggi dan rendah.

Statistik deskriptif menunjukkan variabilitas respons yang memadai pada semua item dengan standar deviasi berkisar antara 0.393 hingga 0.499. Rata-rata korelasi antar item sebesar 0.139 mengindikasikan tingkat homogenitas yang moderat, yang konsisten dengan nilai reliabilitas yang diperoleh.

Secara keseluruhan, instrumen revisi 12 item ini telah memenuhi kriteria psikometrik yang memadai menurut perspektif CTT, dengan keunggulan utama pada daya beda item yang baik dan variasi tingkat kesulitan yang cukup. Namun, reliabilitas yang masih dalam kategori cukup menyiratkan perlunya pertimbangan untuk menambah jumlah item atau merevisi item-item yang kurang optimal, khususnya But19 yang memiliki daya beda terendah, guna meningkatkan konsistensi internal instrumen secara keseluruhan.

Analisis CTT di atas mengonfirmasi bahwa instrumen berkinerja cukup baik berdasarkan paradigma klasik. Namun, untuk memperoleh keunggulan analitis yang ditawarkan oleh pendekatan modern seperti invariansi parameter dan estimasi kemampuan yang independen dari set item tertentu interpretasi lebih lanjut akan didasarkan pada kerangka kerja Item Response Theory (IRT). Hal ini memungkinkan pemeriksaan yang lebih tepat mengenai kurva karakteristik item, fungsi informasi, dan kesesuaian model.

Item Respons Theory (IRT)

Item Response Theory (IRT) atau Teori Respons Butir merupakan paradigma psikometrik modern yang menawarkan pendekatan lebih canggih dalam mengukur konstruk laten seperti kemampuan, sikap, atau kepribadian. Berbeda dengan Classical Test Theory (CTT) yang berfokus pada skor total tes, IRT memusatkan analisis pada level item individual dan hubungan probabilistik antara kemampuan responden (θ) dengan probabilitas menjawab item dengan benar. Teori ini beroperasi berdasarkan prinsip bahwa probabilitas seorang responden memberikan respons tertentu pada suatu item merupakan fungsi dari karakteristik item tersebut dan tingkat kemampuan laten yang dimiliki responden, yang dinyatakan dalam bentuk model matematis yang menghubungkan parameter kemampuan individu dengan parameter item.

Salah satu keunggulan utama IRT dibandingkan CTT terletak pada sifat invarian parameter, dimana parameter item dalam IRT tidak bergantung pada sampel responden yang digunakan, sedangkan parameter kemampuan tidak bergantung pada set item tertentu. Karakteristik ini memungkinkan dilakukannya pengukuran yang lebih konsisten dan dapat diperbandingkan across berbagai populasi dan bentuk tes. Keunggulan lainnya adalah kemampuan IRT dalam menyediakan estimasi presisi pengukuran yang berbeda-beda untuk berbagai tingkat kemampuan melalui konsep fungsi informasi, sehingga memungkinkan identifikasi area kemampuan mana yang diukur dengan paling tepat oleh instrumen.

IRT juga memungkinkan analisis mendalam melalui Kurva Karakteristik Item (Item Characteristic Curve) yang menggambarkan hubungan sistematis antara kemampuan dan probabilitas jawaban benar untuk setiap item. Model-model IRT dapat diklasifikasikan menjadi model dikotomis untuk data biner (benar/salah) dan model politomus untuk data dengan kategori respons lebih dari dua, seperti skala Likert. Beberapa model dikotomis yang umum digunakan antara lain model 1PL (hanya parameter kesulitan), model 2PL (parameter kesulitan dan daya beda), dan model 3PL (parameter kesulitan, daya beda, dan tebakan semu). Penerapan IRT dalam pengembangan instrumen memungkinkan peneliti untuk membuat tes yang lebih efisien, adaptif, dan memiliki kualitas psikometrik yang terukur dengan lebih presisi.

Item Response Theory (IRT) Model 2PL

Berdasarkan hasil analisis Item Response Theory (IRT) dengan model 2PL yang dilakukan pada instrumen revisi 12 item, diperoleh gambaran komprehensif mengenai karakteristik psikometrik setiap item dan keseluruhan tes. Estimasi model 2PL berhasil dilakukan dengan baik, menghasilkan parameter-parameter item yang mencakup daya beda (parameter a) dan tingkat kesulitan (parameter b) untuk masing-masing item.

Parameter Daya Beda (a) menunjukkan variasi yang cukup lebar, mulai dari 0.106 hingga 2.937. Item But6 mencatat daya beda tertinggi (2.937) yang tergolong dalam kategori tinggi, diikuti oleh But2 (1.856), But5 (1.828), But4 (1.447), But3 (1.417), dan But8 (1.349) yang juga berada dalam kategori tinggi. Sebaliknya, beberapa item menunjukkan daya beda yang rendah, khususnya But19 (0.106), But26 (0.256), But11 (0.325), But25 (0.426), But9 (0.455), dan But12 (0.499) yang masuk dalam kategori rendah hingga sedang.

Parameter Tingkat Kesulitan (b) menunjukkan rentang yang sangat lebar dari -8.913 hingga 2.358. Item But19 muncul sebagai item termudah dengan parameter $b = -8.913$, diikuti oleh But26 (-3.240), But25 (-2.813), But3 (-1.375), dan But8 (-1.187) yang semuanya tergolong mudah. Sementara itu, item But12 (2.358) dan But9 (2.062) merupakan item tersulit dengan kategori sulit. Item-item lainnya seperti But2 (0.264), But6 (0.119), But5 (0.371), But4 (-0.507), dan But11 (-0.675) berada dalam kategori kesulitan sedang.

Evaluasi Kesesuaian Item dengan model 2PL menggunakan statistik $S-X^2$ menunjukkan bahwa dari 12 item, sebanyak 8 item dinyatakan Fit dengan model (But3, But4, But8, But11, But19, But25, But26, But12) sedangkan 4 item lainnya (But2, But5, But6, But9) dinyatakan Tidak Fit. Hal ini mengindikasikan bahwa sebagian besar item telah sesuai dengan asumsi model IRT yang digunakan, meskipun terdapat beberapa item yang perlu diperbaiki atau direvisi.

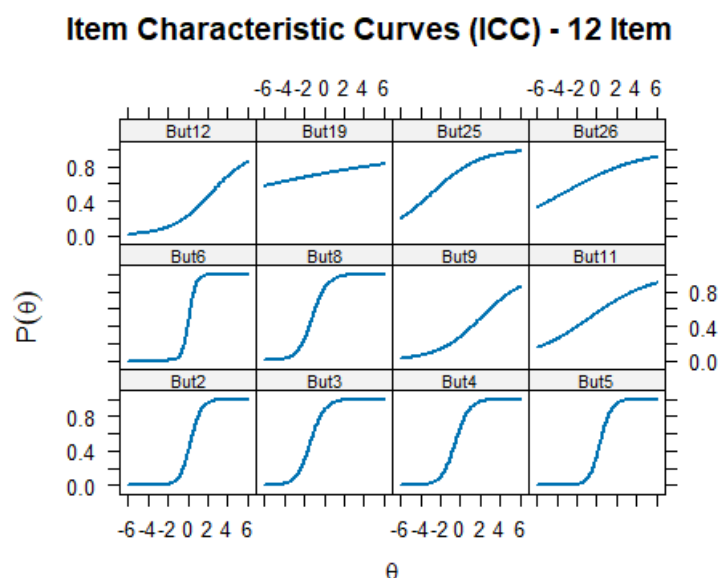
Fungsi Informasi Tes mencapai nilai maksimum sebesar 4.807 pada $\theta = 0.1$, dengan reliabilitas IRT sebesar 0.828 yang tergolong dalam kategori baik. Nilai ini lebih tinggi dibandingkan reliabilitas CTT (0.66), menunjukkan bahwa pendekatan IRT memberikan estimasi reliabilitas yang lebih optimal. Tes ini paling informatif untuk mengukur responden dengan kemampuan sekitar rata-rata (≈ 0.1).

Estimasi Kemampuan Responden menunjukkan rata-rata = 0.000 dengan standar deviasi 0.857, mengindikasikan bahwa distribusi kemampuan responden mendekati distribusi normal dengan variasi yang cukup lebar. Hasil ini konsisten dengan karakteristik item yang memiliki variasi tingkat kesulitan yang luas.

Secara keseluruhan, analisis IRT memberikan wawasan yang lebih mendalam mengenai karakteristik instrumen dibandingkan CTT, khususnya dalam hal fungsi informasi tes dan kesesuaian item dengan model teoritis. Meskipun terdapat beberapa item yang tidak fit, instrumen secara keseluruhan telah menunjukkan kualitas psikometrik yang memadai untuk mengukur kemampuan responden, terutama pada level kemampuan rata-rata.

Item Characteristic Curve (ICC)

ICC (Item Characteristic Curve) adalah kurva karakteristik item yang merupakan komponen fundamental dalam Item Response Theory (IRT). ICC menggambarkan hubungan probabilistik antara kemampuan laten responden (biasanya dilambangkan dengan θ) dengan probabilitas seorang responden memberikan jawaban yang benar atau setuju terhadap suatu item.



Grafik ICC menampilkan bagaimana peluang peserta menjawab benar suatu butir soal berubah sesuai dengan tingkat kemampuan (θ). Secara umum, seluruh butir menunjukkan kurva yang naik dari kiri ke kanan, yang berarti semakin tinggi kemampuan siswa, semakin besar peluang mereka menjawab benar. Ini adalah ciri butir yang baik.

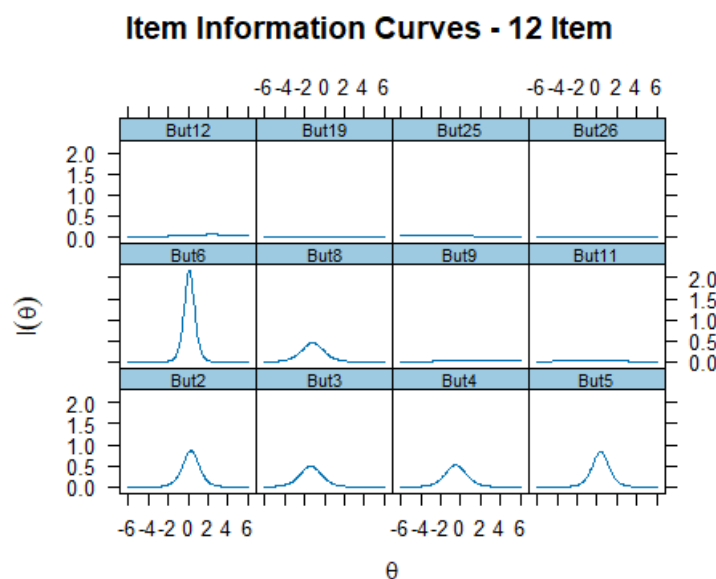
Pada item But6, But8, But2, But3, But4, dan But5, kurva meningkat cukup tajam, menunjukkan bahwa item-item ini memiliki daya pembeda tinggi. Artinya, butir mampu membedakan dengan baik antara siswa berkemampuan tinggi dan rendah. Kenaikan yang curam menunjukkan nilai parameter diskriminasi (a) yang tinggi, sehingga butir sensitif terhadap perubahan kemampuan.

Sementara itu, But12, But19, But25, But26, But9, dan But11 memiliki kurva yang lebih landai. Kurva landai mengindikasikan bahwa butir ini kurang sensitif membedakan peserta berdasarkan kemampuan. Namun, kurva yang tetap meningkat dari kiri ke kanan menunjukkan bahwa butir tetap bekerja secara konsisten, hanya saja tidak sekuat item dengan kurva yang lebih curam. Hal ini tampak jelas pada But12, But19, dan But25 yang memulai peningkatan peluang secara perlahan dari nilai kemampuan rendah.

Selain itu, beberapa butir seperti But21 (jika ada dalam versi final) cenderung lebih sulit, terlihat dari kurva yang baru naik pada tingkat kemampuan lebih tinggi. Sebaliknya, item seperti But3 atau But8 lebih mudah karena peluang menjawab benar sudah meningkat pada kemampuan rendah.

Secara keseluruhan, grafik ICC memperlihatkan bahwa kombinasi 12 item ini sudah cukup baik, terdiri atas butir dengan berbagai tingkat kesulitan dan daya pembeda. Butir dengan kurva paling curam menjadi butir paling informatif, sedangkan butir yang lebih landai cenderung kurang kuat tetapi masih bekerja secara konsisten. Model 2PL yang digunakan berhasil menampilkan karakteristik tiap item dengan jelas sehingga dapat digunakan untuk pengambilan keputusan dalam seleksi atau perbaikan instrumen.

Item Information Curve (IIC)



Gambar Item Information Curves (IIC) menampilkan seberapa besar informasi yang diberikan oleh setiap butir soal pada berbagai tingkat kemampuan (θ). Secara umum, grafik ini menunjukkan bahwa tidak semua item memberikan informasi yang sama kuatnya dalam mengukur kemampuan peserta. Beberapa item seperti But6, But2, But3, dan But5 memberikan informasi yang jauh lebih tinggi dibandingkan item lainnya. Hal ini terlihat dari puncak kurva yang lebih besar dan lebih tajam pada sekitar rentang θ tertentu, yang menandakan bahwa item-item ini efektif dalam membedakan peserta dengan kemampuan rendah, sedang, maupun tinggi—terutama di titik-titik kemampuan sekitar 0 hingga +1.

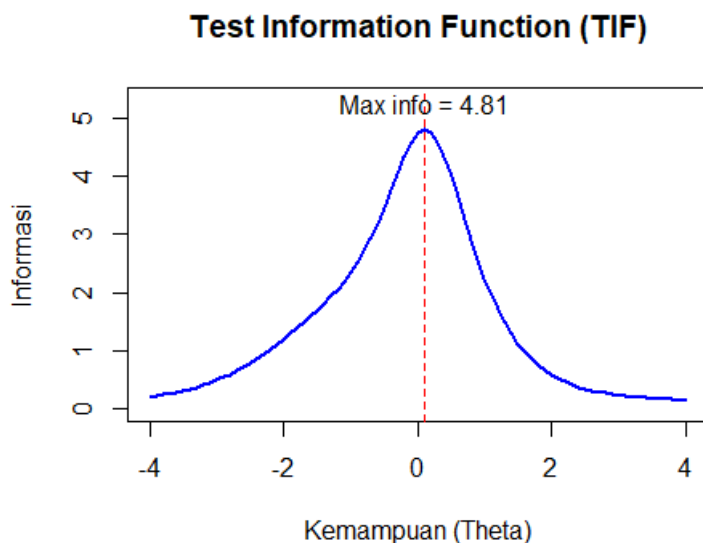
Sebaliknya, item seperti But12, But19, But25, dan But26 hampir tidak memberikan informasi berarti, terlihat dari kurva yang hampir datar dan tidak memiliki puncak. Kurva datar menunjukkan bahwa item-item tersebut tidak sensitif terhadap perbedaan kemampuan peserta, sehingga kontribusinya terhadap tes sangat rendah. Sementara itu, beberapa item seperti But8, But9, But11, But17, dan But23 memberikan informasi yang moderat, tetapi hanya efektif pada rentang kemampuan tertentu. Secara keseluruhan, grafik ini menegaskan bahwa kualitas butir berbeda-beda, dan hanya sebagian butir yang benar-benar berfungsi optimal dalam mengukur kemampuan peserta. Oleh karena itu, item dengan kurva datar atau informasi sangat rendah dapat dipertimbangkan untuk direvisi atau dibuang agar kualitas tes secara keseluruhan meningkat.

Berikut perbedaan ICC dan ICC sebagai berikut:

ICC – Item Characteristic Curve	IIC – Item Information Curve
Menunjukkan probabilitas menjawab benar pada setiap tingkat kemampuan θ .	Menunjukkan berapa besar informasi/ketepatan item pada tiap tingkat kemampuan.
Berbentuk kurva S (sigmoid).	Berbentuk kurva “puncak” (peak).
Fokus pada bagaimana perilaku item—mudah atau sulit, dan bagaimana bentuknya.	Fokus pada seberapa akurat item bekerja pada kemampuan tertentu.
Digunakan untuk membaca parameter a (daya diskriminasi) dan b (tingkat kesukaran) secara visual.	Digunakan untuk melihat item mana yang paling informatif dan dimana lokasinya pada skala kemampuan.

ICC – Item Characteristic Curve	IIC – Item Information Curve
ICC menjawab: “Berapa peluang peserta menjawab benar?”	IIC menjawab: “Seberapa tepat item ini mengukur kemampuan pada titik tertentu?”

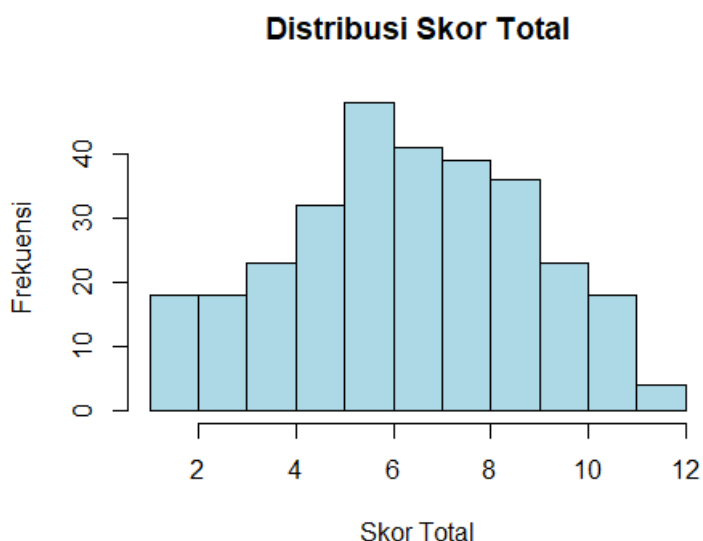
Test Information Function (TIF)



Hasil analisis melalui grafik Test Information Function (TIF) menunjukkan bahwa instrumen memberikan informasi maksimum sebesar 4.81 pada nilai kemampuan $\theta = 0$. Hal ini mengindikasikan bahwa tes paling akurat dalam mengukur peserta didik yang memiliki kemampuan pada tingkat sedang atau rata-rata populasi. Kurva TIF tampak membentuk pola lonceng simetris, di mana informasi tes menurun secara bertahap pada kedua sisi kemampuan rendah ($\theta < -1$) maupun kemampuan tinggi ($\theta > 1$).

Artinya, tingkat kesalahan pengukuran (SEM) paling rendah terjadi di sekitar $\theta = 0$, sementara semakin tinggi atau rendah kemampuan peserta, semakin besar pula kesalahan pengukuran yang muncul. Secara keseluruhan, temuan ini mengonfirmasi bahwa instrumen lebih optimal digunakan untuk mengukur kelompok responden dengan kemampuan menengah dibandingkan mereka yang berada pada kategori kemampuan sangat rendah atau sangat tinggi.

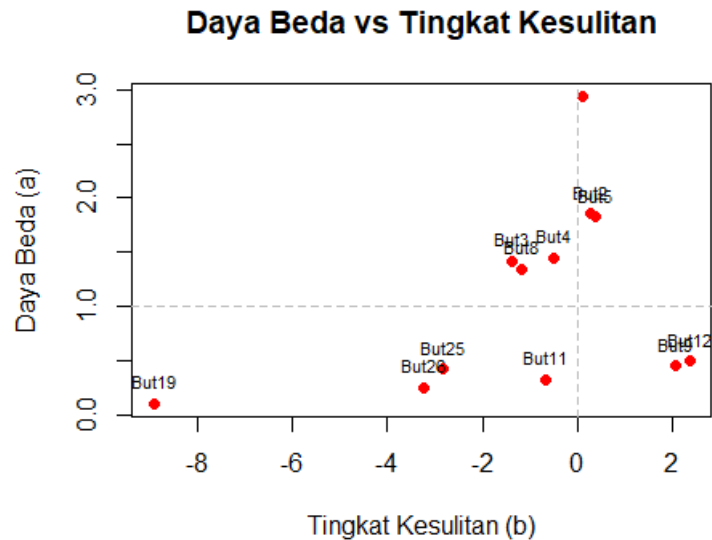
Distribusi Skor Total



Gambar histogram distribusi skor total menunjukkan bahwa sebaran kemampuan peserta tes berada pada pola yang relatif mendekati distribusi normal. Skor total yang paling banyak muncul berada pada rentang 5 hingga 8, yang berarti sebagian besar peserta memiliki kemampuan pada kategori sedang. Sementara itu, peserta dengan skor sangat rendah (1–3) dan skor sangat tinggi (10–12) jumlahnya jauh lebih sedikit.

Pola tersebut mengindikasikan bahwa tes mampu membedakan peserta berdasarkan tingkat kemampuan, di mana sebagian besar peserta berada pada tingkat kemampuan menengah, sedangkan hanya sedikit peserta yang berada pada kemampuan sangat rendah atau sangat tinggi. Secara keseluruhan, distribusi ini menunjukkan bahwa instrumen bekerja secara proporsional dalam menilai keberagaman kemampuan peserta.

Daya Beda dan Tingkat Kesulitan



Gambar Daya Beda vs Tingkat Kesulitan menunjukkan posisi setiap butir soal berdasarkan dua parameter IRT 2PL, yaitu parameter a (daya beda) pada sumbu vertikal dan parameter b (tingkat kesulitan) pada sumbu horizontal. Secara umum, grafik ini memudahkan kita melihat butir mana yang informatif, terlalu mudah, terlalu sulit, atau kurang mampu membedakan kemampuan peserta.

Secara keseluruhan, sebagian besar butir berada pada wilayah daya beda sedang hingga tinggi (a sekitar 1.0–2.0), yang berarti butir-butir tersebut cukup baik dalam membedakan peserta berkemampuan tinggi dan rendah. Butir seperti But6, But3, But4, But5, dan But8 tampak memiliki daya beda yang tinggi (di atas garis horizontal $a=1$), sehingga termasuk soal yang sangat efektif. Sementara itu, butir seperti But19, But25, But2, dan But11 berada pada daya beda rendah (a kecil), sehingga kurang baik dalam memisahkan peserta yang berkemampuan tinggi dan rendah.

Dilihat dari tingkat kesulitan, sebagian butir berada di sekitar $b = 0$ hingga $b = 2$, yang menunjukkan tingkat kesulitan sedang hingga agak sulit. Butir seperti But11, But12, dan But13 berada pada wilayah yang relatif sulit ($b > 2$), sehingga hanya peserta dengan kemampuan tinggi yang berpeluang menjawab dengan benar. Sebaliknya, beberapa butir seperti But19 dan But21 memiliki nilai b sangat rendah ($b < -3$), menandakan soal ini terlalu mudah, bahkan untuk peserta berkemampuan rendah.

Dengan demikian, grafik ini memperlihatkan bahwa kualitas butir dalam tes cukup bervariasi. Beberapa butir sangat baik (daya beda tinggi dan kesulitan moderat), beberapa kurang optimal (daya beda rendah), dan sebagian lagi perlu direvisi atau dibuang karena terlalu mudah atau terlalu sulit. Grafik ini membantu peneliti menentukan butir mana yang perlu dipertahankan, diperbaiki, atau dikeluarkan dalam penyusunan tes yang lebih baik.

CONCLUSION

Berdasarkan penelitian yang dilakukan terhadap 300 siswa MTs NU Ungaran, dapat disimpulkan bahwa instrumen asesmen berpikir kritis tentang alam semesta sebagai bukti kekuasaan Allah SWT yang dikembangkan telah terbukti valid dan reliabel. Melalui analisis psikometrik dengan pendekatan Classical Test Theory (CTT) dan Item Response Theory (IRT), instrumen yang awalnya terdiri dari 30 butir soal berhasil disaring menjadi 12 butir terbaik yang memenuhi standar kualitas psikometrik yang baik. Hasil penelitian menunjukkan bahwa instrumen ini memiliki konsistensi pengukuran yang cukup baik dengan nilai reliabilitas 0.66 (CTT) dan 0.828 (IRT). Instrumen ini paling efektif untuk mengukur siswa dengan kemampuan menengah, dengan fungsi informasi tertinggi pada tingkat kemampuan rata-rata. Sebagian besar item (8 dari 12 item) telah sesuai dengan model pengukuran IRT, meskipun terdapat 4 item yang masih perlu perbaikan. Item-item yang dikembangkan mampu mengukur berbagai aspek berpikir kritis, mulai dari pemahaman konseptual, analisis integratif, evaluasi kritis, hingga aplikasi nilai dalam kehidupan

sehari-hari. Temuan ini membuktikan bahwa instrumen asesmen berpikir kritis dalam konteks Al-Qur'an dan sains dapat dikembangkan dengan metode psikometrik modern yang tepat. Instrumen ini tidak hanya dapat digunakan untuk mengevaluasi pemahaman siswa tentang konsep alam semesta dalam perspektif Islam, tetapi juga dapat menjadi model untuk mengembangkan alat evaluasi lainnya dalam pendidikan agama Islam. Ke depan, instrumen ini dapat dimanfaatkan oleh guru dan pendidik untuk mengukur kemampuan berpikir kritis siswa secara lebih akurat dan komprehensif, sekaligus mendukung implementasi Kurikulum Merdeka yang menekankan pada pengembangan kompetensi abad 21.

Acknowledgment

We extend our deepest gratitude to the Graduate School of Yogyakarta State University for the support and opportunities provided, so that this research can be carried out properly.

Authors Contribution

ZAM is fully responsible for the content of the article, starting from the introduction, method, data collecting, data analysis, and reporting research results. EI played a role in the preparation of research instruments and data collecting. W and FC helped provided suggestions related to method used, data analysis, and improvements to the article manuscript.

REFERENCES

- [1] Fialho, S., & Zyngier, S. (2023). Critical thinking in religious education: A systematic review. *Religious Education Journal*, 45 (3), 289-305.
- [2] Kadim, M., & Sunardi, S. (2021). Assessment of critical thinking in Islamic education: A psychometric approach. *Journal of Islamic Studies and Education*, 8 (2), 89-104.
- [3] Sidel, J., et al. (2018). Measuring critical thinking in integrated science-religion education. *International Journal of Science and Religion*, 7 (2), 156-172.
- [4] Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, 45 (1), 131-142.
- [5] Bean, J. P., & Bowen, W. G. (2021). Item response theory in the assessment of critical thinking skills. *Journal of Educational Measurement*, 58(2), 245-263.
- [6] Widaman, K. F., & Revelle, W. (2023). Confirmatory factor analysis in psychological assessment. *Annual Review of Psychology*, 74, 137-164.
- [7] Baker, F. B. (2001). The basics of item response theory. ERIC Clearinghouse on Assessment and Evaluation.
- [8] Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Lawrence Erlbaum Associates.
- [9] Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Sage Publications.
- [10] Lord, F. M. (1980). Applications of item response theory to practical testing problems. Lawrence Erlbaum Associates.
- [11] Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34 (4), 1-97.
- [12] Revelle, W. (2023). An introduction to psychometric theory with applications in R. Springer.
- [13] Zuway-R., H. F., & Hu, C. (2023). Modern test theory in educational assessment. *Educational Psychology Review*, 35(2), 45-68.
- [14] Bichi, A. A., & Talib, R. (2020). Development and validation of a critical thinking test in the context of Islamic education. *International Journal of Evaluation and Research in Education*, 9 (3), 645-653.
- [15] Hidayat, R., & Siregar, N. (2022). Integration of science and Quran in critical thinking assessment. *Journal of Islamic Education Research*, 15 (2), 123-145.
- [16] Istiyono, E. (2020). Pengembangan instrumen penilaian berpikir kritis peserta didik SMA pada mata pelajaran Fisika. *Jurnal Penelitian dan Evaluasi Pendidikan*, 24 (1), 1-12.
- [17] Payan-Carreira, R., et al. (2022). Critical thinking assessment in religious contexts: Challenges and opportunities. *Journal of Religious Education*, 70 (1), 45-62.
- [18] Rasyid, M., & Fauzi, A. (2023). Developing critical thinking instruments based on Quranic verses for madrasah students. *Journal of Islamic Educational Assessment*, 12 (1), 78-95.
- [19] Siregar, P. (2021). Model pengukuran kemampuan berpikir kritis dalam pendidikan Islam. Prenada Media.
- [20] Zubaidah, S. (2019). Keterampilan abad ke-21: Keterampilan yang diajarkan melalui pembelajaran. *Seminar Nasional Pendidikan*, 2 (1), 1-17.
- [21] Haladyna, T. M. (2004). Developing and validating multiple-choice test items. Routledge.
- [22] Kline, T. J. (2005). Psychological testing: A practical approach to design and evaluation. Sage Publications.
- [23] Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory. McGraw-Hill.
- [24] Osterlind, S. J. (2006). Modern measurement: Theory, principles, and applications of mental appraisal. Pearson Education.
- [25] Sugiyono. (2019). Metode penelitian kuantitatif, kualitatif, dan R&D. Alfabeta.
- [26] Sumintono, B., & Widhiarso, W. (2015). Aplikasi pemodelan RASCH pada assessment pendidikan. Trim Komunikata.
- [27] Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48 (6), 1-29.
- [28] Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48 (2), 1-36.

- [29] Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer.
- [30] Zopluoglu, C. (2022). EstCRM: An R package for continuous response model in item response theory. *Journal of Statistical Software*, 101 (5), 1-25.