E-ISSN: 3031-8521
P-ISSN: 3032-1867

Halaman 3403-3418
Volume 3 Nomor 4 Tahun 2026

Pengembangan Instrumen *Two Tier Multiple Choice* Berbasis Indikator Ennis untuk Mengukur Kemampuan Berpikir Kritis Pada Materi Struktur Atom Menggunakan Model *Rasch*

Najjmi Nur'aini[1], Okta Suryani[2]

Program Studi Pendidikan Kimia Departemen Kimia Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Negeri Padang
e-mail : najjminuraini@gmail.com[1], okta.suryani.os@fmipa.unp.ac.id[2]

Abstrak

Kemampuan berpikir kritis merupakan kompetensi penting dalam pembelajaran kimia, tetapi instrumen pilihan ganda konvensional belum sepenuhnya mengungkap alasan di balik jawaban murid. Penelitian ini bertujuan mengembangkan instrumen Two Tier Multiple Choice berbasis indikator Ennis untuk mengukur kemampuan berpikir kritis pada materi struktur atom. Penelitian ini menggunakan metode Research and Development (R&D) dengan analisis model Rasch. Populasi penelitian adalah murid kelas X Fase E SMAN 15 Padang tahun ajaran 2026/2027, dengan sampel sebanyak 60 murid. Instrumen penelitian meliputi soal Two Tier Multiple Choice, lembar validasi ahli, dan pedoman penskoran. Data dianalisis menggunakan Content Validity Index (CVI), Content Validity Ratio (CVR), serta statistik kesesuaian dan reliabilitas model Rasch melalui Ministep. Hasil penelitian menunjukkan nilai CVI sebesar 0,980 sampai 1,000, CVR sebesar 0,6 sampai 1,0, person reliability 0,92, item reliability 0,99, dan Cronbach's Alpha 0,87. Namun, butir S2 dan S3 menunjukkan ketidaksesuaian dengan model Rasch. Kesimpulannya, instrumen memiliki validitas isi dan konsistensi pengukuran yang baik, tetapi beberapa butir masih perlu ditinjau sebelum digunakan secara lebih luas.

Kata kunci: *Asesmen Sains, Berpikir Kritis, Model Rasch, Struktur Atom, Tes Pilihan Ganda Dua Tingkat*

Abstract

Critical thinking is an essential competency in chemistry learning, but conventional multiple-choice tests do not fully reveal the reasoning underlying students' answers. This study aimed to develop a Two-Tier Multiple Choice instrument based on Ennis's indicators to measure critical thinking skills in atomic structure. The study employed a Research and Development (R&D) method with Rasch model analysis. The population comprised Grade X students at SMAN 15 Padang in the 2026/2027 academic year, with

60 students selected as the sample. The research instruments included Two-Tier Multiple Choice questions, an expert validation form, and a scoring rubric. Data were analyzed using the Content Validity Index (CVI), Content Validity Ratio (CVR), and Rasch model fit and reliability statistics through Ministep. The results showed a CVI of 0.980 to 1.000, a CVR of 0.6 to 1.0, person reliability of 0.92, item reliability of 0.99, and Cronbach's Alpha of 0.87. However, items S2 and S3 exhibited misfit with the Rasch model. In conclusion, the instrument demonstrated strong content validity and measurement consistency, although several items require further review before broader implementation.

Keywords: *Critical Thinking, Rasch Model, Science Assessment, Structure of Atom, Two-Tier Multiple Choice*

LATAR BELAKANG

Kemampuan berpikir kritis merupakan salah satu kemampuan esensial yang perlu dikuasai murid dalam proses pembelajaran karena mendukung pemahaman konsep, penyelesaian masalah, dan penerapan pengetahuan dalam kehidupan nyata (Ariadila dkk., 2023). Dalam pembelajaran kimia, kemampuan berpikir kritis memiliki peranan penting karena karakteristik konsep kimia yang kompleks dan abstrak menuntut murid untuk memahami fenomena melalui representasi makroskopik, submikroskopik, dan simbolik (Budi Prasetya, 2017). Murid tidak hanya dituntut untuk menjelaskan konsep, tetapi juga menghubungkan ketiga representasi tersebut secara tepat (Fadillah, 2025). Namun, sebagian besar murid masih menganggap kimia sebagai mata pelajaran yang sulit dipelajari karena konsep-konsepnya saling berkaitan dan berjenjang serta membutuhkan pemahaman konseptual yang kuat (Langitasari et al., 2021).

Salah satu materi kimia yang membutuhkan pemahaman konseptual dan kemampuan berpikir kritis adalah struktur atom. Materi ini mencakup perkembangan teori atom, struktur atom, jumlah proton, elektron, dan neutron, nomor atom, nomor massa, isotop, isobar, isoton, serta massa atom relatif yang tidak dapat diamati secara langsung. Karakteristik tersebut menuntut murid untuk menganalisis hubungan antarpartikel dan memahami keterkaitan konsep yang bersifat abstrak. Oleh karena itu, pemahaman materi struktur atom tidak hanya memerlukan penguasaan konsep, tetapi juga kemampuan berpikir analitis dan kritis serta kerja sama antarmurid dalam memahami materi (Haleda et al., 2025). Kondisi ini menunjukkan pentingnya instrumen penilaian yang mampu mengukur kemampuan berpikir kritis murid secara lebih mendalam, khususnya dalam memahami materi struktur atom.

Permasalahan yang muncul dalam pengukuran kemampuan berpikir kritis adalah keterbatasan instrumen penilaian yang digunakan dalam pembelajaran. Soal pilihan ganda konvensional hanya mengukur ketepatan jawaban tanpa mengungkap alasan atau proses berpikir yang mendasarinya, sehingga sulit membedakan murid yang benar-benar memahami konsep dengan murid yang menjawab benar karena menebak. Berdasarkan hasil observasi awal dan wawancara dengan guru kimia di SMAN 15 Padang, instrumen penilaian yang digunakan masih didominasi oleh soal pilihan ganda

konvensional dan soal uraian sederhana yang belum mampu mengukur kemampuan berpikir kritis murid secara optimal (Fadillah, 2025). Kondisi tersebut menunjukkan perlunya instrumen penilaian yang tidak hanya memperhatikan jawaban akhir, tetapi juga menggali alasan murid dalam menjawab pertanyaan.

Salah satu alternatif yang dapat digunakan untuk mengatasi keterbatasan tersebut adalah instrumen Two Tier Multiple Choice (TTMC). Instrumen ini dikembangkan oleh Treagust pada 1988 sebagai alat diagnostik untuk mengidentifikasi pemahaman konsep dan miskonsepsi murid. TTMC terdiri atas dua bagian, yaitu bagian pertama yang memuat pertanyaan dengan beberapa pilihan jawaban dan bagian kedua yang meminta murid menyatakan alasan atas jawaban yang dipilih (Bayrak, 2013). Melalui kedua bagian tersebut, penilaian tidak hanya berfokus pada ketepatan jawaban, tetapi juga mempertimbangkan alasan yang mendasarinya. Dengan demikian, TTMC berpotensi digunakan untuk memperoleh informasi yang lebih mendalam mengenai pemahaman konsep dan proses berpikir murid. Pengembangan instrumen ini perlu didasarkan pada indikator berpikir kritis Ennis agar pengukuran kemampuan berpikir kritis pada materi struktur atom memiliki landasan yang jelas.

Selain pengembangan bentuk instrumen, diperlukan analisis untuk mengetahui karakteristik dan kualitas pengukurannya. Instrumen yang dikembangkan perlu diuji berdasarkan validitas isi, validitas empiris, reliabilitas person dan item, tingkat kesukaran butir, daya beda berdasarkan Point Measure Correlation (PTMEA Corr.), serta kesesuaian butir dengan model Rasch. Distribusi kemampuan murid juga perlu dianalisis melalui Wright Map untuk memperoleh gambaran mengenai sebaran kemampuan murid dan tingkat kesulitan butir. Oleh karena itu, penggunaan model Rasch dalam penelitian ini diperlukan untuk menganalisis kualitas empiris instrumen yang dikembangkan. Kajian mengenai pengembangan instrumen TTMC berbasis indikator berpikir kritis Ennis pada materi struktur atom dengan analisis model Rasch masih terbatas berdasarkan kajian penelitian terdahulu yang diuraikan, sehingga diperlukan penelitian pengembangan yang mengintegrasikan ketiga aspek tersebut.

Penelitian ini bertujuan menghasilkan instrumen Two Tier Multiple Choice berbasis indikator berpikir kritis Ennis untuk mengukur kemampuan berpikir kritis murid pada materi struktur atom serta mengetahui karakteristik dan kualitas instrumen berdasarkan validitas isi, validitas empiris, reliabilitas person dan item, tingkat kesukaran butir, daya beda, kesesuaian butir dengan model Rasch, dan distribusi kemampuan murid melalui Wright Map. Penelitian ini penting dilakukan karena instrumen penilaian yang mampu mengungkap jawaban sekaligus alasan murid diperlukan untuk memperoleh gambaran kemampuan berpikir kritis secara lebih mendalam. Kebaruan penelitian ini terletak pada pengembangan instrumen TTMC yang mengintegrasikan indikator berpikir kritis Ennis dengan materi struktur atom dan analisis model Rasch untuk mengevaluasi karakteristik serta kualitas empiris instrumen. Hasil penelitian diharapkan dapat menjadi alternatif alat asesmen bagi pendidik, sarana latihan bagi murid, dan referensi bagi peneliti selanjutnya dalam pengembangan instrumen pengukuran kemampuan berpikir kritis.

METODE

Penelitian ini menggunakan metode penelitian dan pengembangan (Research and Development atau R&D) yang bertujuan menghasilkan instrumen Two Tier Multiple Choice berbasis indikator berpikir kritis Ennis untuk mengukur kemampuan berpikir kritis murid pada materi struktur atom. Pengembangan instrumen dilakukan dengan memperhatikan kesesuaian butir soal dengan indikator yang diukur serta kualitas instrumen berdasarkan hasil validasi ahli dan pengujian empiris. Model Rasch digunakan sebagai pendekatan analisis untuk mengevaluasi karakteristik pengukuran instrumen melalui respons murid terhadap butir soal yang dikembangkan. Model ini memungkinkan analisis kemampuan murid dan tingkat kesulitan butir dalam satu skala logit sehingga kualitas instrumen dapat dievaluasi secara objektif (Boone et al., 2014; Sumintono & Widhiarso, 2015).

Instrumen pengumpulan data dalam penelitian ini meliputi lembar wawancara guru, kuesioner validasi ahli, instrumen Two Tier Multiple Choice, kisi-kisi instrumen, dan pedoman penskoran. Lembar wawancara digunakan untuk memperoleh informasi mengenai kemampuan berpikir kritis murid dan bentuk instrumen penilaian yang digunakan di SMAN 15 Padang. Kuesioner validasi digunakan untuk menilai kelayakan instrumen yang dikembangkan, sedangkan instrumen Two Tier Multiple Choice digunakan untuk mengukur kemampuan berpikir kritis murid melalui dua bagian, yaitu pilihan jawaban pada tier pertama dan pilihan alasan pada tier kedua. Kisi-kisi instrumen disusun berdasarkan keterkaitan antara tujuan pembelajaran, materi struktur atom, indikator dan subindikator berpikir kritis Ennis, stimulus, jawaban, serta alasan. Pedoman penskoran digunakan untuk menentukan skor berdasarkan kombinasi jawaban dan alasan murid.

Data penelitian terdiri atas data kualitatif dan kuantitatif. Data kualitatif diperoleh dari wawancara guru serta saran dan komentar validator ahli, sedangkan data kuantitatif diperoleh dari hasil validasi ahli dan respons murid terhadap instrumen yang dikembangkan. Validasi isi dilakukan oleh lima orang ahli menggunakan Content Validity Index (CVI) berdasarkan nilai Content Validity Ratio (CVR). Validitas empiris dianalisis menggunakan model Rasch dengan bantuan perangkat lunak Ministep untuk mengevaluasi kesesuaian butir, reliabilitas, tingkat kesulitan, daya beda, dan kemampuan murid. Analisis kesesuaian butir dilakukan dengan memperhatikan nilai Outfit Mean Square (MNSQ), Infit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), dan Point Measure Correlation (PTMEA Corr.). Kriteria kesesuaian butir mengacu pada rentang Outfit MNSQ dan Infit MNSQ sebesar 0,5 sampai 1,5, ZSTD sebesar -2 sampai +2, serta PTMEA Corr. sebesar 0,4 sampai 0,85 (Sumintono & Widhiarso, 2013; Azizah & Wahyuningsih, 2020).

Analisis reliabilitas dilakukan dengan menginterpretasikan nilai Person Reliability dan Item Reliability untuk mengetahui konsistensi respons murid dan kestabilan kualitas butir. Selain itu, nilai separation index dianalisis untuk mengetahui kemampuan instrumen dalam membedakan tingkat kemampuan murid dan tingkat kesulitan butir. Daya beda butir dievaluasi berdasarkan nilai PTMEA Corr. dengan mempertimbangkan arah dan kekuatan hubungan antara respons pada setiap butir dan kemampuan murid.

Sementara itu, tingkat kesulitan butir dianalisis berdasarkan nilai measure logit dan simpangan baku logit serta ditinjau melalui Wright Map, yang menggambarkan distribusi kemampuan murid dan kesulitan butir pada skala yang sama (Sumintono & Widhiarso, 2015). Hasil analisis tersebut digunakan untuk mengevaluasi karakteristik empiris instrumen dan menentukan butir yang perlu dipertahankan atau ditinjau kembali.

Penelitian dilaksanakan di SMA Negeri 15 Padang dan Departemen Kimia FMIPA Universitas Negeri Padang pada tahun akademik 2026/2027. Subjek penelitian terdiri atas validator ahli dan responden uji coba instrumen. Validator ahli berjumlah lima orang dosen kimia dari FMIPA Universitas Negeri Padang yang memiliki kompetensi dalam bidang kimia, evaluasi pembelajaran, atau pengembangan instrumen. Responden uji coba terdiri atas 60 murid kelas X Fase E SMAN 15 Padang yang telah mempelajari materi struktur atom. Pemilihan responden dilakukan secara sengaja dengan mempertimbangkan kesesuaian karakteristik responden dengan tujuan pengembangan instrumen. Objek penelitian ini adalah instrumen Two Tier Multiple Choice berbasis indikator berpikir kritis Ennis pada materi struktur atom.

Prosedur pengembangan instrumen mengacu pada tahapan pengembangan instrumen berbasis model Rasch yang dikemukakan oleh Liu (2010). Tahap pertama adalah menetapkan tujuan pengukuran dan populasi sasaran, yaitu mengembangkan instrumen untuk mengukur kemampuan berpikir kritis murid kelas X pada materi struktur atom. Tahap kedua adalah mendefinisikan konstruk kemampuan berpikir kritis berdasarkan landasan teori yang digunakan. Tahap ketiga adalah mengidentifikasi dimensi dan indikator konstruk, kemudian menyusun kisi-kisi instrumen yang memuat indikator, subindikator, materi, stimulus, butir soal, pilihan jawaban, dan pilihan alasan. Instrumen yang telah disusun selanjutnya divalidasi oleh lima orang ahli untuk menilai kesesuaian aspek materi, konstruksi, dan bahasa. Hasil validasi digunakan sebagai dasar untuk mengevaluasi dan memperbaiki instrumen sebelum diujicobakan kepada murid.

Tahap berikutnya adalah melakukan uji coba instrumen kepada 60 murid kelas X Fase E SMAN 15 Padang yang telah mempelajari materi struktur atom. Respons murid diberi skor berdasarkan kombinasi jawaban pada kedua tier. Skor 3 diberikan apabila jawaban pada tier pertama dan alasan pada tier kedua benar. Skor 2 diberikan apabila jawaban pada tier pertama benar, tetapi alasan pada tier kedua salah. Skor 1 diberikan apabila jawaban pada tier pertama salah, tetapi alasan pada tier kedua benar. Skor 0 diberikan apabila jawaban dan alasan keduanya salah atau tidak diisi. Data hasil uji coba selanjutnya dianalisis menggunakan model Rasch melalui perangkat lunak Ministep untuk memperoleh informasi mengenai kemampuan murid, tingkat kesulitan butir, kesesuaian butir, daya beda, reliabilitas, dan distribusi kemampuan melalui Wright Map. Hasil analisis digunakan untuk meninjau kesesuaian butir dengan model, mengevaluasi kualitas instrumen, serta menentukan kebutuhan revisi terhadap butir yang belum memenuhi kriteria. Tahap akhir meliputi penetapan klaim validitas dan penyusunan panduan penggunaan instrumen berdasarkan hasil pengembangan dan analisis yang telah dilakukan.

HASIL DAN PEMBAHASAN

Hasil Pengembangan dan Validasi Instrumen

Instrumen yang dikembangkan berupa Two Tier Multiple Choice yang terdiri atas dua tingkat, yaitu pilihan jawaban (tier 1) dan pilihan alasan (tier 2). Pengembangan instrumen mengacu pada indikator berpikir kritis Ennis yang disesuaikan dengan materi struktur atom dan disusun berdasarkan kisi-kisi. Instrumen terdiri atas 10 butir soal, yaitu S1 sampai S10. Validasi isi dilakukan oleh lima validator untuk menilai kesesuaian indikator, materi, konstruksi, dan penyajian soal. Hasil validasi dianalisis menggunakan CVI dan CVR sebagai dasar evaluasi dan perbaikan butir sebelum pengujian empiris.

Tabel 1. Validasi Isi Instrumen

Aspek	Hasil
Jumlah validator	5 orang
Jumlah butir yang dinilai	10 butir
Metode analisis	CVI dan CVR
Nilai CVI	0,980-1,000
Nilai CVR setiap butir	0,6-1,0
Keputusan akhir validasi	Esensial

Hasil validasi isi menunjukkan bahwa nilai Content Validity Index (CVI) berada pada rentang 0,980–1,000. Hasil tersebut menunjukkan bahwa instrumen memperoleh penilaian relevansi isi yang tinggi dari para validator. Sementara itu, nilai Content Validity Ratio (CVR) pada setiap butir bervariasi, dengan sebagian besar butir memperoleh nilai 1,0 dan beberapa butir memperoleh nilai 0,6. Variasi tersebut menunjukkan adanya perbedaan tingkat kesepakatan validator dalam menilai esensialitas butir. Oleh karena itu, hasil CVI dan CVR perlu dipertimbangkan bersama-sama dengan hasil analisis Rasch untuk mengevaluasi kualitas instrumen. Validasi isi memberikan bukti mengenai relevansi dan esensialitas butir, sedangkan analisis Rasch memberikan informasi mengenai karakteristik butir berdasarkan pola respons responden.

Hasil Analisis Menggunakan Model *Rasch*

Analisis model *Rasch* dilakukan menggunakan perangkat lunak *Ministep* untuk mengevaluasi karakteristik butir dan respons responden. Analisis mencakup reliabilitas responden dan butir, tingkat kesulitan butir, kesesuaian butir dengan model, serta distribusi kemampuan responden terhadap tingkat kesulitan soal.

Reliabilitas Responden dan Butir Soal

Analisis data dalam penelitian ini dilakukan menggunakan model *Rasch* berbantuan perangkat lunak *Ministep* versi 5.12.1. Berdasarkan hasil *summary statistics*, analisis melibatkan 60 murid dan 10 butir soal. Hasil analisis reliabilitas responden dan butir soal serta konsistensi internal instrumen disajikan pada Gambar 1.

SUMMARY OF 59 MEASURED (NON-EXTREME) Person

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	17.8	10.0	1.49	.70	.92	-.13	.60	.02
SEM	.8	.0	.38	.02	.10	.14	.17	.18
P.SD	5.8	.0	2.86	.18	.74	1.08	1.26	1.36
S.SD	5.9	.0	2.89	.18	.75	1.09	1.28	1.37
MAX.	29.0	10.0	6.54	1.15	3.12	2.21	9.90	9.91
MIN.	4.0	10.0	-4.81	.52	.10	-2.24	.08	-.79

REAL RMSE .80 TRUE SD 2.75 SEPARATION 3.45 Person RELIABILITY .92
MODEL RMSE .72 TRUE SD 2.77 SEPARATION 3.83 Person RELIABILITY .94
S.E. OF Person MEAN = .38

MAXIMUM EXTREME SCORE: 1 Person 1.7%

SUMMARY OF 60 MEASURED (EXTREME AND NON-EXTREME) Person

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	18.1	10.0	1.60	.72				
SEM	.8	.0	.38	.03				
P.SD	6.0	.0	2.95	.23				
S.SD	6.0	.0	2.98	.24				
MAX.	30.0	10.0	7.85	1.86				
MIN.	4.0	10.0	-4.81	.52				

REAL RMSE .83 TRUE SD 2.83 SEPARATION 3.43 Person RELIABILITY .92
MODEL RMSE .76 TRUE SD 2.85 SEPARATION 3.77 Person RELIABILITY .93
S.E. OF Person MEAN = .38

Person RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .87 SEM = 2.14
STANDARDIZED (50 ITEM) RELIABILITY = .99

SUMMARY OF 10 MEASURED (NON-EXTREME) Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	108.3	60.0	.00	.30	.93	-.27	2.31	1.58
SEM	19.4	.0	1.26	.03	.11	.43	1.27	1.36
P.SD	58.2	.0	3.77	.08	.32	1.28	3.81	4.07
S.SD	61.3	.0	3.97	.08	.34	1.35	4.01	4.30
MAX.	173.0	60.0	4.46	.45	1.50	1.89	9.90	9.91
MIN.	36.0	60.0	-4.79	.21	.44	-2.81	.11	-.91

REAL RMSE .32 TRUE SD 3.75 SEPARATION 11.70 Item RELIABILITY .99
MODEL RMSE .31 TRUE SD 3.75 SEPARATION 12.26 Item RELIABILITY .99
S.E. OF Item MEAN = 1.26

Item RAW SCORE-TO-MEASURE CORRELATION = -1.00
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000

Gambar 1. *Summary Statistics*

Sumber: *Output ministep, summary statistic*

Hasil analisis menunjukkan nilai *person reliability* sebesar 0,92, *person separation* sebesar 3,43, *item reliability* sebesar 0,99, *item separation* sebesar 11,70, dan *Cronbach's Alpha* (KR-20) sebesar 0,87.

Nilai *person reliability* sebesar 0,92 menunjukkan konsistensi pengukuran yang tinggi dalam membedakan kemampuan murid berdasarkan pola responsnya. Hal ini didukung oleh nilai *person separation* sebesar 3,43 yang menunjukkan bahwa instrumen mampu membedakan responden ke dalam beberapa tingkat kemampuan. Sementara itu, nilai *item reliability* sebesar 0,99 menunjukkan bahwa hierarki tingkat kesulitan butir memiliki kestabilan yang sangat tinggi berdasarkan sampel yang dianalisis. Nilai *item separation* sebesar 11,70 juga menunjukkan bahwa butir soal memiliki perbedaan tingkat kesulitan yang dapat dibedakan secara jelas.

Selain itu, nilai *Cronbach's Alpha* (KR-20) sebesar 0,87 menunjukkan konsistensi internal respons yang baik. Secara keseluruhan, hasil tersebut menunjukkan bahwa instrumen memiliki karakteristik reliabilitas yang baik berdasarkan data penelitian. Namun, reliabilitas yang tinggi belum cukup untuk menyimpulkan bahwa seluruh butir telah memenuhi kriteria kesesuaian model *Rasch* (Linacre, 2002). Oleh karena itu, hasil ini perlu ditinjau bersama analisis tingkat kesulitan (*item measure*) dan kesesuaian butir (*item fit*) untuk memperoleh gambaran yang lebih menyeluruh mengenai kualitas instrumen dalam mengukur kemampuan berpikir kritis pada materi struktur atom.

Tingkat Kesukaran Butir Soal (Item Measure)

Analisis instrumen Two Tier Multiple Choice berbasis indikator berpikir kritis Ennis pada materi struktur atom dilakukan menggunakan model Rasch melalui perangkat lunak Ministep versi 5.12.1. Instrumen terdiri atas 10 butir soal yang diujikan kepada 60 peserta didik Fase E SMA dengan penskoran empat kategori, yaitu 3 jika kedua tier benar, 2 jika hanya tier 1 benar, 1 jika hanya tier 2 benar, dan 0 jika keduanya salah atau tidak dijawab sesuai ketentuan penelitian (Yamtinah et al., 2015). Butir soal disusun berdasarkan indikator berpikir kritis Ennis, meliputi fokus pada pertanyaan, analisis argumen, klarifikasi, penilaian hasil pengamatan, penarikan kesimpulan, evaluasi definisi, serta pengambilan keputusan (Rintayati et al., 2021). Analisis model Rasch juga digunakan untuk menentukan tingkat kesulitan butir (item measure) berdasarkan pola respons peserta didik. Nilai measure yang lebih tinggi menunjukkan butir relatif lebih sulit, sedangkan nilai yang lebih rendah menunjukkan butir relatif lebih mudah dijawab.

Person: REAL SEP.: 3.43 REL.: .92 ... Item: REAL SEP.: 11.70 REL.: .99

Item STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
5	36	60	4.46	.24	.44	-2.81	.26	-.91	.71	.67	89.8	75.0	S5
10	43	60	4.08	.23	1.50	1.89	1.09	.39	.69	.70	71.2	69.8	S10
2	48	60	3.83	.22	.66	-1.58	9.90	9.48	.71	.72	61.0	65.6	S2
9	57	60	3.40	.21	.93	-.24	.60	-.37	.78	.75	71.2	66.6	S9
3	70	60	2.84	.21	1.06	.37	9.90	9.91	.74	.79	66.1	63.0	S3
6	161	60	-3.10	.33	1.02	.18	.31	-.47	.70	.69	86.4	84.7	S6
7	163	60	-3.32	.34	.64	-1.09	.18	-.78	.68	.67	88.1	84.3	S7
8	164	60	-3.44	.35	1.10	.40	.32	-.45	.64	.65	88.1	85.7	S8
4	168	60	-3.96	.37	1.34	.97	.40	-.26	.58	.60	88.1	88.3	S4
1	173	60	-4.79	.45	.62	-.83	.11	-.78	.52	.50	93.2	92.0	S1
MEAN	108.3	60.0	.00	.30	.93	-.27	2.31	1.58			80.3	77.5	
P.SD	58.2	.0	3.77	.08	.32	1.28	3.81	4.07			11.0	10.1	

Gambar 2. *Item Measure*

Sumber: *Output ministep, Item measure*

Berdasarkan hasil analisis *Ministep*, butir S5 memiliki nilai measure tertinggi sebesar 4,46 logit, sedangkan S1 memiliki nilai terendah sebesar −4,79 logit. Hasil tersebut menunjukkan bahwa S5 merupakan butir yang paling sulit dan S1 merupakan butir yang paling mudah bagi kelompok responden dalam penelitian ini. Perbedaan nilai tersebut memperlihatkan adanya variasi tingkat kesulitan di antara butir soal yang dikembangkan.

Variasi tingkat kesulitan memberikan informasi mengenai posisi relatif setiap butir dalam instrumen. Namun, tingkat kesulitan tidak dapat dijadikan satu-satunya dasar untuk menilai kualitas butir atau menentukan tingkat kemampuan berpikir kritis yang

diukur. Oleh karena itu, hasil *item measure* perlu ditafsirkan bersama hasil kesesuaian butir (*item fit*) serta kesesuaian isi soal dengan indikator berpikir kritis Ennis.

Kesesuaian Butir Soal (Item Fit)

Analisis *item fit* dilakukan untuk mengevaluasi kesesuaian pola respons murid dengan prediksi model *Rasch*. Statistik yang digunakan meliputi *Infit Mean Square* (MNSQ), *Outfit Mean Square* (MNSQ), dan *Point Measure Correlation* (PTMEA Corr.). Dalam penelitian ini, rentang MNSQ sebesar 0,5–1,5 digunakan sebagai kriteria umum untuk mengidentifikasi butir yang produktif dalam pengukuran. Nilai MNSQ yang terlalu tinggi menunjukkan pola respons yang tidak sesuai dengan prediksi model (*underfit*), sedangkan nilai yang terlalu rendah menunjukkan pola respons yang lebih terprediksi daripada yang diharapkan (*overfit*) (Linacre, 2002).

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
2	48	60	3.83	.22	.66	-1.58	9.90	9.48	A .71	.72	61.0	65.6	S2
3	70	60	2.84	.21	1.06	.37	9.90	9.91	B .74	.79	66.1	63.0	S3
10	43	60	4.08	.23	1.50	1.89	1.09	.39	C .69	.70	71.2	69.8	S10
4	168	60	-3.96	.37	1.34	.97	.40	-.26	D .58	.60	88.1	88.3	S4
8	164	60	-3.44	.35	1.10	.40	.32	-.45	E .64	.65	88.1	85.7	S8
6	161	60	-3.10	.33	1.02	.18	.31	-.47	e .70	.69	86.4	84.7	S6
9	57	60	3.40	.21	.93	-.24	.60	-.37	d .78	.75	71.2	66.6	S9
7	163	60	-3.32	.34	.64	-1.09	.18	-.78	c .68	.67	88.1	84.3	S7
1	173	60	-4.79	.45	.62	-.83	.11	-.78	b .52	.50	93.2	92.0	S1
5	36	60	4.46	.24	.44	-2.81	.26	-.91	a .71	.67	89.8	75.0	S5
MEAN	108.3	60.0	.00	.30	.93	-.27	2.31	1.58			80.3	77.5	
P.SD	58.2	.0	3.77	.08	.32	1.28	3.81	4.07			11.0	10.1	

Gambar 3. *Item Fit Order*

Sumber: *Output ministep, Item fit order*

Hasil analisis item fit menunjukkan bahwa butir S2 dan S3 memiliki nilai Outfit MNSQ sebesar 9,90, melebihi batas 2,0 sehingga pola responsnya belum sesuai dengan prediksi model Rasch. Meskipun keduanya memiliki nilai PTMEA Corr. positif, respons murid pada kedua butir perlu ditelaah lebih lanjut untuk mengidentifikasi kemungkinan sumber penyimpangan. Butir S10 memiliki Infit MNSQ sebesar 1,50 dan Outfit MNSQ sebesar 1,09, sedangkan S9 menunjukkan nilai masing-masing 0,93 dan 0,60. Butir S1, S4, S5, S6, S7, dan S8 cenderung mengalami overfit, sehingga perlu dievaluasi bersama statistik lainnya. Seluruh butir memiliki nilai PTMEA Corr. positif pada rentang 0,52–0,78, yang menunjukkan hubungan searah dengan kemampuan berpikir kritis yang diukur (Linacre, 1994). Secara keseluruhan, evaluasi butir perlu mempertimbangkan nilai Infit MNSQ, Outfit MNSQ, PTMEA Corr., serta kesesuaian isi soal dengan indikator berpikir kritis sebelum menentukan keputusan untuk mempertahankan, merevisi, atau menghapus butir.

Analisis Pola Respons pada Butir S2 dan S3

Analisis Pola Respons Butir S2

Butir S2 memiliki nilai measure 3,83 logit sehingga tergolong relatif sulit. Butir ini mengukur kemampuan menganalisis argumen melalui identifikasi bukti eksperimen Rutherford yang menyanggah model atom Thomson beserta alasannya. Hasil analisis Rasch menunjukkan nilai Infit MNSQ 0,66, Outfit MNSQ 9,90, dan PTMEA Corr. 0,71. Nilai Outfit MNSQ yang tinggi menunjukkan pola respons yang tidak sesuai dengan

prediksi model, sedangkan PTMEA Corr. positif menunjukkan hubungan searah dengan konstruk yang diukur.

Pola respons skor 1, yaitu jawaban tier 1 salah tetapi alasan tier 2 benar, menunjukkan ketidaksesuaian antara jawaban dan alasan. Oleh karena itu, evaluasi S2 perlu memperhatikan kejelasan stimulus, kesesuaian jawaban dan alasan, serta keterkaitannya dengan indikator menganalisis argumen. Namun, penyebab tingginya nilai Outfit MNSQ belum dapat dipastikan hanya berdasarkan hasil Item Fit Order.

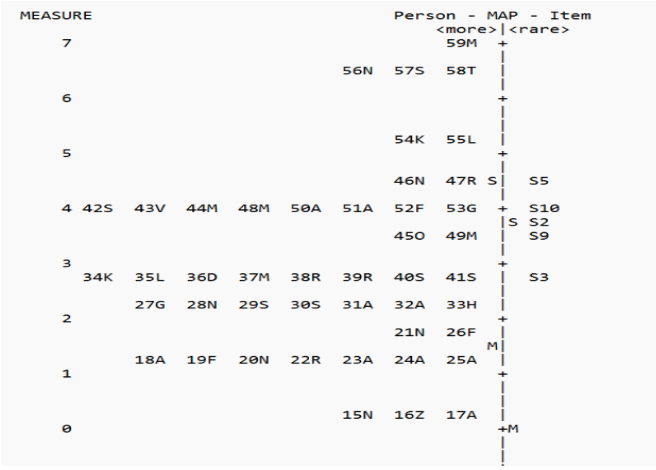
Analisis Pola Respons Butir S3

Butir S3 memiliki nilai measure 2,84 logit sehingga tergolong relatif sulit. Butir ini mengukur kemampuan mengajukan dan menjawab pertanyaan klarifikasi melalui evaluasi klaim tentang pemantulan partikel alfa oleh inti atom. Hasil analisis Rasch menunjukkan nilai Infit MNSQ 1,06, Outfit MNSQ 9,90, dan PTMEA Corr. 0,74. Nilai Infit MNSQ memenuhi kriteria, sedangkan Outfit MNSQ yang tinggi mengindikasikan pola respons yang tidak sesuai dengan prediksi model. PTMEA Corr. positif menunjukkan hubungan searah dengan kemampuan yang diukur.

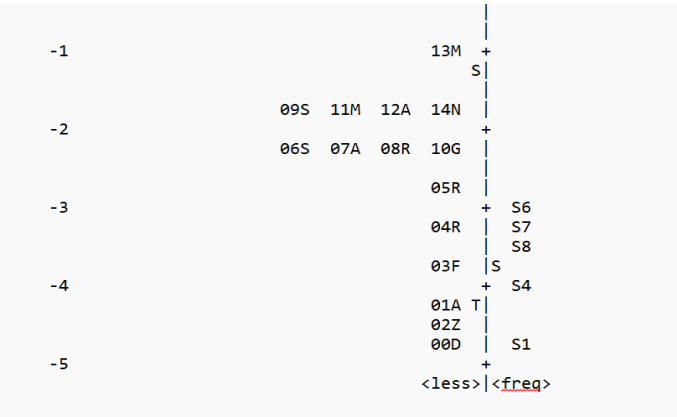
Pola respons skor 2, yaitu jawaban tier pertama benar tetapi alasan tier kedua salah, menunjukkan ketidakkonsistenan antara jawaban dan alasan. Oleh karena itu, evaluasi S3 perlu memperhatikan kejelasan stimulus, kesesuaian pilihan jawaban dan alasan, serta keterkaitannya dengan indikator kemampuan berpikir kritis. Namun, penyebab ketidaksesuaian pola respons belum dapat dipastikan hanya berdasarkan hasil Item Fit Order dan memerlukan evaluasi lebih lanjut.

Analisis Peta *Wright (Wright Map)*

Wright Map digunakan untuk menggambarkan distribusi kemampuan murid dan tingkat kesulitan butir soal pada skala logit yang sama. Peta ini membantu menunjukkan posisi relatif kemampuan murid terhadap kesulitan butir, sehingga dapat digunakan untuk menilai sejauh mana rentang kesulitan instrumen mencakup kemampuan responden.



E-ISSN: 3031-8521
P-ISSN: 3032-1867
Halaman 3403-3418
Volume 3 Nomor 4 Tahun 2026



Gambar 4. *Wright Map*
Sumber: *Output ministep, Wright Map*

Hasil analisis Wright Map menunjukkan bahwa estimasi kemampuan murid berada pada rentang −4,81 hingga 7,85 logit, sedangkan tingkat kesulitan butir berada pada rentang −4,79 hingga 4,46 logit. Murid 59M memiliki kemampuan tertinggi sebesar 7,85 logit, sedangkan S5 merupakan butir tersulit dengan nilai measure 4,46 logit. Sebaliknya, kemampuan terendah murid sebesar −4,81 logit dan S1 merupakan butir termudah dengan nilai measure −4,79 logit.

Variasi tingkat kesulitan butir, seperti S3 sebesar 2,84 logit dan S6 sebesar −3,10 logit, menunjukkan bahwa instrumen mencakup tingkat kesulitan yang beragam. Namun, kemampuan murid 59M berada 3,39 logit di atas butir tersulit, sehingga mengindikasikan kemungkinan terbatasnya tantangan bagi murid berkemampuan tinggi. Secara keseluruhan, Wright Map dapat menjadi dasar evaluasi cakupan kesulitan instrumen. Penambahan butir yang lebih sulit dapat dipertimbangkan dengan memperhatikan distribusi kemampuan seluruh murid dan kesesuaian butir terhadap rentang kemampuan tersebut.

Analisis Butir Berdasarkan Indikator Berpikir Kritis Ennis

Analisis butir berdasarkan indikator berpikir kritis Ennis dilakukan dengan menelaah kisi-kisi, isi soal, pilihan jawaban dan alasan, serta hasil analisis model Rasch. Butir S2 mengukur kemampuan menganalisis argumen melalui penelaahan pernyataan berdasarkan informasi yang tersedia, sedangkan S3 mengukur kemampuan mengajukan dan menjawab pertanyaan klarifikasi atau tantangan terkait materi struktur atom. Meskipun keduanya termasuk indikator penjelasan dasar, tuntutan berpikir yang diukur berbeda sesuai karakteristik masing-masing butir.

Tabel 1. Pemetaan Butir Soal dan Indikator Berpikir Kritis Ennis

Butir	Sub indikator	Hasil Analisis *Rasch*
S1	Focus pada suatu pertanyaan	*Measure*: -4,79 Logit *Infit* MNSQ: 0,62 *Outfit* MNSQ: 0,11 PT. Measur Corr: 0,52

S2	Analisis argumen	*Measure*: 3,83 Logit *Infit* MNSQ: 0,66 *Outfit* MNSQ: 9,90 PT. Measur Corr: 0,71
S3	Ajukan dan menjawab pertanyaan klarifikasi atau tantangan	*Measure*: 2,84logit *Infit* MNSQ: 1,06 *Outfit* MNSQ: 9,90 PT. Measur Corr: 0,74
S4	Focus pada suatu pertanyaan	*Measure*: -3,96 Logit *Infit* MNSQ: 1,34 *Outfit* MNSQ: 0,40 PT. Measur Corr: 0,58
S5	Analisis argumen	*Measure*: 4,46 Logit *Infit* MNSQ: 0,44 *Outfit* MNSQ: 0,26 PT. Measur Corr: 0,71
S6	Amati dan menilai laporan hasil pengamatan	*Measure*: -3.10 Logit *Infit* MNSQ: 1,02 *Outfit* MNSQ: 0,31 PT. Measur Corr: 0,70
S7	Membuat dan mempertimbangkan kesimpulan	*Measure*: -3,32logit *Infit* MNSQ: 0,64 *Outfit* MNSQ: 0,18 PT. Measur Corr: 0,68
S8	Mendefinisikan istilah dan menilai definisi	*Measure*: -3,44 Logit *Infit* MNSQ: 1,10 *Outfit* MNSQ: 0,32 PT. Measur Corr: 0,64
S9	Mengintegrasikan disposisi dan kemampuan lain dalam membuat dan mempertahankan keputusan	*Measure*: 3,40 Logit *Infit* MNSQ: 0,93 *Outfit* MNSQ: 0,60 PT. Measur Corr: 0,78
S10	Membuat dan mempertimbangkan kesimpulan	*Measure*: 4.08 Logit *Infit* MNSQ: 1,50 *Outfit* MNSQ: 1,09 PT. Measur Corr: 0,69

Hasil analisis *Rasch* menunjukkan bahwa S2 dan S3 memiliki nilai *Outfit* MNSQ sebesar 9,90. Temuan ini menjadi dasar untuk menelaah kembali hubungan antara isi soal, alasan yang tersedia, dan respons responden. Namun, hasil tersebut tidak secara langsung membuktikan bahwa indikator Ennis yang digunakan tidak tepat. Penilaian terhadap keterwakilan indikator tetap perlu didasarkan pada kesesuaian isi soal dengan kisi-kisi dan tujuan pengukuran. Pemetaan tersebut menjadi dasar untuk menghubungkan hasil pengukuran dengan karakteristik setiap indikator. Analisis terhadap seluruh butir perlu dilakukan secara konsisten agar dapat diketahui apakah setiap soal telah merepresentasikan kemampuan yang direncanakan dalam pengembangan instrumen.

Pembahasan

Kualitas Pengukuran Instrumen Berdasarkan Model Rasch

Hasil analisis menunjukkan bahwa instrumen memiliki nilai *person reliability* sebesar 0,92, *item reliability* sebesar 0,99, dan *Cronbach's alpha* sebesar 0,87. Secara umum, hasil tersebut menunjukkan konsistensi pengukuran dan kestabilan hierarki tingkat kesulitan butir yang baik pada data penelitian yang dianalisis.

Nilai *person reliability* memberikan informasi mengenai konsistensi pembedaan kemampuan responden, sedangkan *item reliability* menggambarkan kestabilan hierarki butir berdasarkan sampel yang digunakan. Adapun *Cronbach's alpha* memberikan gambaran mengenai konsistensi internal respons secara keseluruhan. Ketiga statistik tersebut memiliki fungsi yang berbeda sehingga interpretasinya tidak seharusnya disamakan.

Meskipun demikian, hasil reliabilitas perlu dipertimbangkan bersama statistik kesesuaian butir. Keberadaan butir S2 dan S3 dengan nilai *Outfit* MNSQ sebesar 9,90 menunjukkan bahwa masih terdapat aspek yang perlu ditelaah lebih lanjut. Dengan demikian, reliabilitas yang tinggi merupakan salah satu informasi mengenai kualitas pengukuran, tetapi bukan satu-satunya dasar untuk menyatakan seluruh butir telah memenuhi karakteristik model Rasch.

Karakteristik Tingkat Kesulitan dan Kesesuaian Butir

Hasil *item measure* menunjukkan bahwa S1 memiliki tingkat kesulitan terendah sebesar −4,79 logit, sedangkan S5 memiliki tingkat kesulitan tertinggi sebesar 4,46 logit. Perbedaan tersebut menunjukkan adanya variasi tingkat kesulitan dalam instrumen.

Variasi tingkat kesulitan perlu dipahami berdasarkan karakteristik kelompok responden dan tuntutan yang terdapat pada setiap soal. Butir yang lebih sulit tidak selalu menuntut kemampuan berpikir kritis yang lebih tinggi, karena kesulitan juga dapat dipengaruhi oleh penguasaan konsep, kejelasan redaksi, dan karakteristik pilihan jawaban serta alasan.

Selain tingkat kesulitan, hasil *item fit* menunjukkan bahwa S2 dan S3 memerlukan penelaahan lebih lanjut. Nilai *Outfit* MNSQ sebesar 9,90 pada kedua butir mengindikasikan adanya pola respons yang tidak sesuai dengan prediksi model. Temuan ini perlu dikaji melalui pemeriksaan isi soal dan distribusi respons, bukan hanya melalui interpretasi statistik.

Karena instrumen terdiri atas dua *tier*, pemeriksaan juga perlu memperhatikan kesesuaian antara jawaban *tier* pertama dan alasan *tier* kedua. Meskipun demikian, format dua *tier* tidak dapat langsung dinyatakan sebagai penyebab ketidaksesuaian. Kesimpulan mengenai penyebabnya memerlukan bukti dari karakteristik butir dan data respons yang tersedia.

Keterwakilan Indikator Berpikir Kritis Ennis

Keterwakilan indikator berpikir kritis Ennis dalam instrumen ditentukan melalui kesesuaian antara kisi-kisi, tujuan pengukuran, dan isi setiap butir. Penggunaan indikator sebagai dasar penyusunan soal membantu memastikan bahwa butir dikembangkan untuk mengukur kemampuan yang telah direncanakan.

Perbedaan karakteristik S2 dan S3 memperlihatkan bahwa butir yang berada dalam kelompok indikator yang sama dapat memiliki sasaran kemampuan yang berbeda. S2 berfokus pada kemampuan menganalisis argumen, sedangkan S3 berfokus pada kemampuan mengajukan dan menjawab pertanyaan klarifikasi atau tantangan. Perbedaan tersebut perlu tercermin dalam tuntutan berpikir, konteks permasalahan, dan alasan yang disediakan pada setiap butir.

Hasil analisis *Rasch* memberikan informasi empiris mengenai karakteristik respons terhadap butir, tetapi tidak dengan sendirinya membuktikan bahwa seluruh indikator berpikir kritis telah terwakili secara memadai. Oleh karena itu, keterwakilan indikator perlu dinilai bersama hasil validasi isi dan pemeriksaan substansi soal.

Dengan demikian, evaluasi instrumen perlu membedakan antara keterwakilan indikator secara konseptual dalam rancangan soal dan kualitas pengukuran yang ditunjukkan oleh data empiris. Kedua aspek tersebut saling melengkapi dalam menilai kualitas instrumen yang dikembangkan.

Implikasi Hasil Analisis terhadap Perbaikan Instrumen

Hasil penelitian menunjukkan bahwa instrumen *two tier multiple choice* yang dikembangkan memiliki konsistensi pengukuran yang baik berdasarkan nilai reliabilitas yang diperoleh. Selain itu, hasil analisis tingkat kesulitan menunjukkan adanya variasi karakteristik antarbutir, dengan S5 sebagai butir tersulit dan S1 sebagai butir termudah. Temuan ini menunjukkan bahwa butir soal memiliki tingkat kesulitan yang berbeda dalam mengukur kemampuan berpikir kritis responden.

Meskipun demikian, hasil analisis *item fit* menunjukkan bahwa S2 dan S3 perlu ditelaah lebih lanjut karena nilai *Outfit* MNSQ kedua butir mencapai 9,90. Penelaahan perlu difokuskan pada kesesuaian antara indikator berpikir kritis, materi struktur atom, redaksi soal, pilihan jawaban pada tingkat pertama, dan alasan pada tingkat kedua. Hal ini penting karena instrumen *two tier multiple choice* tidak hanya menilai ketepatan jawaban, tetapi juga mempertimbangkan alasan yang mendasari jawaban tersebut.

Evaluasi terhadap S2 dan S3 sebaiknya tidak hanya didasarkan pada nilai Outfit MNSQ, tetapi juga mempertimbangkan hasil validasi isi, nilai Infit MNSQ, PTMEA Corr., serta kesesuaian butir dengan indikator berpikir kritis yang diukur. Dengan demikian, keputusan untuk mempertahankan, merevisi, atau menguji kembali butir perlu didasarkan pada keseluruhan bukti yang diperoleh dalam penelitian. Hasil evaluasi tersebut dapat menjadi dasar untuk menyempurnakan instrumen agar karakteristik setiap butir semakin sesuai dengan tujuan pengukuran.

SIMPULAN

Penelitian ini menghasilkan instrumen Two Tier Multiple Choice berbasis indikator berpikir kritis Ennis pada materi struktur atom yang terdiri atas 10 butir soal dan diuji kepada 60 murid kelas X Fase E SMAN 15 Padang. Hasil validasi isi oleh lima validator menunjukkan nilai Content Validity Index (CVI) sebesar 0,980 sampai 1,000 dan Content Validity Ratio (CVR) sebesar 0,6 sampai 1,0. Analisis model Rasch menggunakan Ministep menunjukkan nilai person reliability sebesar 0,92, item reliability sebesar 0,99, person separation sebesar 3,43, item separation sebesar 11,70, dan Cronbach’s Alpha

sebesar 0,87. Tingkat kesulitan butir bervariasi, dengan S1 sebagai butir termudah sebesar -4,79 logit dan S5 sebagai butir tersulit sebesar 4,46 logit. Namun, butir S2 dan S3 menunjukkan nilai Outfit Mean Square (MNSQ) sebesar 9,90, sehingga pola responsnya belum sesuai dengan prediksi model Rasch dan memerlukan penelaahan lebih lanjut. Temuan ini menunjukkan bahwa instrumen memiliki konsistensi pengukuran yang baik, tetapi kualitas setiap butir tetap perlu dievaluasi berdasarkan kesesuaian model dan substansi soal.

Penelitian ini memiliki keterbatasan pada jumlah responden yang terbatas pada 60 murid dari satu sekolah, sehingga hasilnya belum dapat digeneralisasikan secara luas. Selain itu, masih terdapat butir yang menunjukkan ketidaksesuaian dengan model Rasch, sementara penyebabnya belum dapat dipastikan hanya berdasarkan statistik pengukuran. Penelitian selanjutnya disarankan melibatkan responden dari lebih banyak sekolah, menguji instrumen pada sampel yang lebih beragam, serta menelaah kembali redaksi soal, pilihan jawaban, dan alasan pada butir yang belum sesuai. Pengujian lanjutan juga diperlukan untuk memperkuat bukti validitas dan kestabilan karakteristik instrumen. Secara praktis, instrumen ini dapat menjadi alternatif asesmen bagi guru kimia untuk mengidentifikasi kemampuan berpikir kritis murid melalui jawaban dan alasan yang diberikan. Hasil analisisnya juga dapat digunakan sebagai dasar untuk memperbaiki butir soal dan merancang evaluasi pembelajaran struktur atom yang lebih sesuai dengan kemampuan murid.

DAFTAR PUSTAKA

Ariadila, S. N., Silalahi, Y. F. N., Fadiyah, F. H., Jamaludin, U., & Setiawan, S. (2023). Analisis pentingnya keterampilan berpikir kritis terhadap pembelajaran bagi siswa. 9(20), 664–669.

Azizah, & Wahyuningsih, S. (2020). Penggunaan model Rasch untuk analisis instrumen the use of Rasch model for analyzing test. Jurnal Pendidikan Matematika (JUPITEK), 3(1), 45–50.

Bayrak, B. K. (2013). Using two-tier test to identify primary students' conceptual understanding and alternative conceptions in acid base. Mevlana International Journal of Education, 3(2), 19–26. https://doi.org/10.13054/mije.13.21.3.2

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Rasch analysis in the human sciences. Springer.

Budi Prasetya, F. (2017). Pengaruh representasi mikroskopik dinamik dan statik melalui strategi REACT terhadap hasil belajar dan motivasi belajar mahasiswa pada materi elektrokimia. Jurnal Pembelajaran Sains, 1(2), 26–33. http://journal2.um.ac.id/index.php/

Ennis, R. H. (1996). Critical thinking dispositions: Their nature and assessability. Informal Logic, 18(2), 165–182.

Fadillah, A. (2025). Struktur atom menggunakan instrumen tes four-tier multiple choice.

Haleda, P., Isa, I., & Munandar, H. (2025). Pengaruh model pembelajaran project based learning terhadap kemampuan berpikir kritis siswa pada materi struktur atom-nanoteknologi Padria. 20, 29–36.

Langitasari, I., Rogayah, T., & Solfarina. (2021). Problem-based learning (PBL) pada topik struktur atom: Keaktifan, kreativitas dan prestasi belajar siswa. 15(2).

Liu, X. (2010). Using and developing measurement instruments in science education. SAGE Publications.

Linacre, J. M. (1994). Sample size and item calibration or person measure stability. Rasch Measurement Transactions, 7(4), 328. http://www.rasch.org/rmt/rmt74m.htm

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? Rasch Measurement Transactions.

Rintayati, P., Lukitasari, H., & Syawaludin, A. (2021). Development of two-tier multiple choice test to assess Indonesian elementary students' higher-order thinking skills. 14(1), 555–566.

Sumintono, B., & Widhiarso, W. (2013). Aplikasi model Rasch untuk penelitian ilmu-ilmu sosial. Trim Komunikata.

Sumintono, B., & Widhiarso, W. (2015). Aplikasi pemodelan Rasch pada assessment pendidikan. Trim Komunikata.

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. International Journal of Science Education. https://doi.org/10.1080/0950069880100204

Yamtinah, S., Saputro, S., & Haryono, H. (2015). Instrumen alternatif untuk penilaian keterampilan proses sains (KPS) dan berfungsi diagnostik pada aspek pengetahuan. Jurnal Materi dan Pembelajaran Fisika, 5(2). http://jurnal.fkip.uns.ac.id/index.php/fisika/article/view/7781