

Evaluasi Psikometrik Butir Instrumen Literasi Sains pada Materi Momentum dan Impuls Berdasarkan Model IRT Tiga Parameter Logistik (3PL)

Della Apriyani Kusuma Putri^{*1}, Ina Ana Khoeriah², Muhammad Firdaus³

^{1,2} Pendidikan IPA, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Muhammadiyah Bogor Raya, Indonesia

³ Pendidikan Kepelatihan Olahraga, Fakultas Kesehatan dan Olahraga, Universitas Muhammadiyah Bogor Raya, Indonesia

Email: 1akpdella@gmail.com, 2inaanakhoeriah@gmail.com, 3firdausumbarapko@gmail.com

Abstrak

Penelitian ini bertujuan mengkarakterisasi butir instrumen literasi sains pada materi momentum dan impuls menggunakan pendekatan Item Response Theory (IRT) model tiga parameter logistik (3PL). Penelitian dilakukan dengan metode research and development (R&D) pada tahap uji coba terbatas terhadap 114 peserta didik SMA yang telah mempelajari materi momentum dan impuls. Instrumen berupa tes pilihan ganda yang dikembangkan berdasarkan indikator literasi sains yang mencakup kemampuan menjelaskan fenomena ilmiah, menafsirkan data, dan mengevaluasi isu berbasis sains. Uji asumsi unidimensionalitas dilakukan melalui analisis faktor eksploratori (EFA), yang menunjukkan satu faktor dominan (KMO = 0,781; RMSEA = 0,063). Analisis IRT 3PL menggunakan jMetrik menghasilkan estimasi parameter diskriminasi (a), kesulitan (b), dan peluang menebak (c). Hasil menunjukkan sebagian besar butir berada pada kategori moderat hingga baik, namun terdapat beberapa butir dengan diskriminasi rendah dan satu butir dengan diskriminasi negatif, tingkat kesulitan sangat ekstrem, serta nilai guessing relatif tinggi. Fungsi informasi tes mengindikasikan presisi pengukuran tertinggi pada rentang kemampuan sedang. Secara umum, instrumen menunjukkan potensi yang baik, tetapi memerlukan revisi pada beberapa butir sebelum digunakan pada skala luas. Penelitian ini memberikan kontribusi dalam pengembangan instrumen literasi sains berbasis analisis IRT untuk meningkatkan validitas dan presisi pengukuran.

Kata kunci: *analisis IRT 3PL, karakterisasi butir, literasi sains, momentum dan impuls, validasi instrumen.*

Abstract

Response Theory (IRT) model. The research employed a research and development (R&D) design at the limited pilot-testing stage involving 114 senior high school students who had studied momentum and impulse. The instrument consisted of multiple-choice items developed based on scientific literacy indicators, including explaining scientific phenomena, interpreting data and evidence, and evaluating science-based issues. The assumption of unidimensionality was examined using Exploratory Factor Analysis (EFA), which indicated a dominant single factor (KMO = 0.781; RMSEA = 0.063). The 3PL IRT analysis was conducted using jMetrik to estimate item discrimination (a), difficulty (b), and guessing (c) parameters. The results showed that most items fell within moderate to good categories; however, several items demonstrated low discrimination, one item showed negative discrimination, some items exhibited extremely high difficulty levels, and several had relatively high guessing parameters. The Test Information Function indicated that the instrument provided the highest measurement precision at the medium ability level. Overall, the instrument shows good potential but requires revision of several items before being implemented on a larger scale. This study contributes to the development of scientific literacy instruments using IRT analysis to enhance measurement validity and precision.

Keywords: *IRT 3PL analysis, item characterization, momentum and impulse, scientific literacy, instrument validation.*

1. PENDAHULUAN

Literasi sains merupakan kompetensi esensial dalam pendidikan abad ke-21 karena mencakup kemampuan individu untuk memahami konsep ilmiah, mengevaluasi proses penyelidikan, serta menafsirkan data secara kritis dalam pengambilan keputusan. Kerangka literasi sains modern menekankan integrasi pengetahuan konten, prosedural, dan epistemik dalam konteks kehidupan nyata (*PISA 2022 Assessment and Analytical Framework*, 2023). Namun demikian, berbagai laporan nasional dan internasional menunjukkan bahwa capaian literasi sains peserta didik Indonesia masih berada pada kategori rendah hingga sedang (Yusmar & Fadilah, 2023), termasuk pada materi fisika yang menuntut pemahaman konseptual dan kemampuan analitis seperti momentum dan impuls (Adianto & Rusli, 2021). Kondisi ini menunjukkan urgensi pengembangan instrumen asesmen yang mampu mengukur kemampuan literasi sains secara lebih akurat dan diagnostik.

Dalam konteks evaluasi pembelajaran, kualitas instrumen sangat menentukan ketepatan interpretasi kemampuan peserta didik. Penelitian terdahulu mengenai literasi sains umumnya masih menggunakan pendekatan Classical Test Theory (CTT) yang bergantung pada karakteristik sampel dan belum sepenuhnya mengevaluasi parameter butir secara mendalam (Rusilowati et al., 2024). Sementara itu, studi yang secara khusus menerapkan analisis Item Response Theory (IRT) model tiga parameter logistik (3PL) pada topik momentum dan impuls masih terbatas. Kesenjangan ini menunjukkan perlunya penelitian yang tidak hanya mengembangkan instrumen, tetapi juga mengevaluasi kualitas psikometriknya secara komprehensif menggunakan pendekatan IRT.

Secara teoretis, model IRT terdiri atas beberapa bentuk, yaitu model satu parameter logistik (1PL), dua parameter logistik (2PL), dan tiga parameter logistik (3PL). Model 1PL hanya mempertimbangkan tingkat kesukaran, sedangkan 2PL menambahkan parameter daya pembeda. Model 3PL memberikan perluasan dengan memasukkan parameter peluang menebak (*guessing*), yang relevan pada tes pilihan ganda (Ariyadi, 2025). Dalam konteks literasi sains berbentuk pilihan ganda, kemungkinan respons benar karena tebakan tidak dapat diabaikan, sehingga model 3PL dipandang lebih representatif dibandingkan 1PL dan 2PL. Secara konseptual, penggunaan 3PL juga berdampak pada desain distraktor, karena efektivitas distraktor berpengaruh terhadap estimasi parameter tebakan.

Selain itu, analisis berbasis IRT memungkinkan estimasi parameter butir yang relatif independen dari sampel serta penyediaan informasi fungsi tes pada berbagai tingkat kemampuan. Penggunaan perangkat lunak seperti jMetrik mendukung estimasi parameter diskriminasi (*a*), kesukaran (*b*), dan tebakan (*c*) secara kuantitatif serta visualisasi kurva karakteristik butir untuk interpretasi yang lebih komprehensif (Aksu et al., 2019).

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengevaluasi kualitas psikometrik butir instrumen literasi sains pada materi momentum dan impuls menggunakan Model IRT Tiga Parameter Logistik (3PL). Penelitian ini diharapkan dapat memperkuat landasan empiris pengembangan instrumen asesmen fisika serta memberikan kontribusi terhadap peningkatan kualitas evaluasi literasi sains di sekolah.

2. METODE PENELITIAN

Penelitian ini menggunakan metode *research and development* (R&D) yang difokuskan pada tahap pengembangan dan uji coba terbatas (*pilot testing*) sebelum implementasi skala luas. Tahap pilot bertujuan memperoleh estimasi awal parameter butir dan mengevaluasi kualitas instrumen sebelum dilakukan kalibrasi final (Knekta et al., 2020). Tujuan penelitian adalah mengkarakterisasi butir instrumen literasi sains pada materi momentum dan impuls menggunakan Item Response Theory (IRT) model tiga parameter logistik (3PL).

2.1. Desain Penelitian

1. Analisis kebutuhan dan kajian literatur dilakukan untuk mengidentifikasi urgensi pengembangan instrumen serta landasan teoritis literasi sains dan materi momentum–impuls.

2. Penyusunan kisi-kisi berbasis indikator literasi sains (Gormally et al., 2012)), bertujuan memastikan setiap butir soal merepresentasikan konstruk yang diukur, seperti kemampuan menjelaskan fenomena, menganalisis data, dan mengevaluasi bukti ilmiah.
3. Pengembangan butir pilihan ganda dilakukan berdasarkan kisi-kisi yang telah disusun dengan memperhatikan aspek konseptual dan konteks aplikatif.
4. Validasi isi oleh ahli bertujuan menilai kesesuaian materi, konstruksi soal, dan kejelasan bahasa.
5. Revisi instrumen dilakukan berdasarkan saran validator untuk meningkatkan kualitas butir.
6. Uji coba terbatas bertujuan memperoleh data empiris awal untuk melihat karakteristik respons peserta didik.
7. Analisis butir menggunakan IRT 3PL dilakukan untuk mengestimasi parameter diskriminasi (a), kesukaran (b), dan tebakan (c), serta mengevaluasi kualitas butir sebelum digunakan pada skala yang lebih luas.

Sebelum estimasi parameter dilakukan, diuji asumsi unidimensionalitas menggunakan analisis faktor eksploratori. Asumsi ini penting karena model IRT mengharuskan pengukuran satu konstruk dominan (Gómez-Benito et al., 2018)

2.2 Subjek Penelitian

Subjek penelitian adalah peserta didik sekolah menengah yang telah mempelajari materi momentum dan impuls. Penelitian ini merupakan uji coba terbatas (pilot study), peserta yang menjadi responden pada penelitian ini berjumlah 114 orang. Dalam model 3PL, estimasi parameter stabil umumnya membutuhkan sampel >200–300 responden, khususnya untuk estimasi parameter tebakan (c) (Hooper et al., 2025). Oleh karena itu, hasil estimasi pada tahap ini diposisikan sebagai estimasi awal (*preliminary calibration*) yang akan dikonfirmasi kembali pada tahap uji skala luas.

2.3 Instrumen Penelitian

Instrumen penelitian berupa tes literasi sains pilihan ganda dengan skor dikotomus (1 = benar; 0 = salah). Pengembangan butir mengacu pada kerangka literasi sains (Gormally et al., 2012) yang meliputi kemampuan menjelaskan fenomena ilmiah, menafsirkan data dan bukti ilmiah, serta mengevaluasi isu berbasis sains. Butir dirancang menggunakan konteks autentik untuk mendukung validitas konstruk, sejalan dengan kerangka asesmen literasi sains PISA 2022 yang menekankan penerapan konsep dalam situasi nyata (*PISA 2022 Results (Volume I)*, 2023).

2.4 Uji Asumsi Unidimensional

Sebelum estimasi parameter dilakukan, instrumen diuji asumsi unidimensionalitas menggunakan analisis faktor eksploratori (EFA) (Ocy et al., 2025). Hasil analisis menunjukkan nilai Kaiser-Meyer-Olkin (KMO) sebesar 0,781, yang berada pada kategori baik, serta uji Bartlett's Test of Sphericity menunjukkan $\chi^2 = 861$, $df = 300$, $p < 0,001$, sehingga data dinyatakan layak untuk analisis faktor.

Berdasarkan hasil ekstraksi, faktor pertama memiliki eigenvalue sebesar 5,702 dan menjelaskan 22,8% varians, sedangkan faktor kedua memiliki eigenvalue 1,037 dan selanjutnya berada di bawah 1. Pola scree plot menunjukkan satu faktor dominan dengan penurunan tajam setelah faktor pertama.

Indeks kecocokan model juga menunjukkan hasil yang memadai dengan nilai RMSEA = 0,063 (90% CI = 0,049–0,078), yang berada dalam kategori fit yang dapat diterima (Mannan et al., 2025).

Temuan tersebut mengindikasikan bahwa instrumen cenderung memenuhi asumsi unidimensionalitas sebagai prasyarat analisis Item Response Theory (IRT) model tiga parameter logistik (3PL), meskipun beberapa butir dengan nilai MSA rendah tetap dipertimbangkan untuk evaluasi lebih lanjut.

2.5 Teknik Analisis Data

Butir soal dinyatakan layak apabila memenuhi kriteria parameter berdasarkan model Item Response Theory (IRT) tiga parameter logistik (3PL). Parameter diskriminasi (a) ditetapkan minimal sebesar 0,30, yang menunjukkan bahwa butir memiliki kemampuan membedakan peserta didik berkemampuan tinggi dan rendah secara memadai (Baker, 2001). Parameter kesukaran (b) berada pada rentang -3 hingga $+3$, yang mencerminkan bahwa tingkat kesukaran butir masih berada dalam distribusi kemampuan yang wajar pada skala kemampuan normal. Parameter tebakan (c) ditetapkan kurang dari 0,25, sesuai dengan probabilitas tebak acak pada soal pilihan ganda dengan empat opsi jawaban.

Selain memenuhi kriteria parameter tersebut, butir juga harus menunjukkan kecocokan terhadap model melalui indeks item fit. Butir yang tidak memenuhi kriteria parameter maupun indeks kecocokan model dipertimbangkan untuk direvisi atau dieliminasi sebelum digunakan pada tahap kalibrasi skala luas.

Evaluasi kualitas tes secara keseluruhan dilakukan melalui analisis Test Information Function (TIF) untuk mengetahui tingkat presisi pengukuran pada berbagai rentang kemampuan peserta didik. Fungsi informasi ini memberikan gambaran sejauh mana instrumen mampu mengukur kemampuan secara akurat pada level tertentu (Kalkan et al., 2020).

3. HASIL DAN PEMBAHASAN

3.1. Parameter Daya Pembeda (a)

Daya pembeda adalah kemampuan suatu butir soal untuk membedakan peserta didik yang memiliki kemampuan tinggi dan kemampuan rendah. Dengan kata lain, butir soal yang baik harus dapat dijawab benar oleh peserta berkemampuan tinggi dan dijawab salah oleh peserta berkemampuan rendah.

Tabel 1. Parameter 1 PL

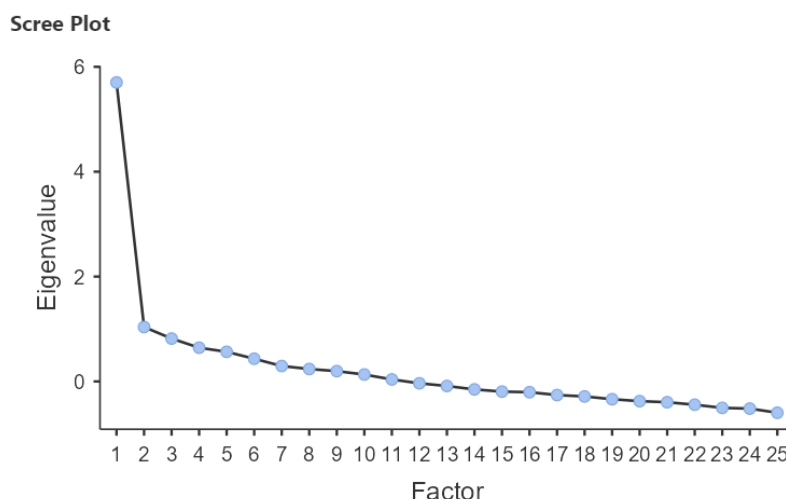
Item	Code	Apar	(SE)
1	L3	1.77	(0.61)
2	L3	1.77	(0.53)
3	L3	1.42	(0.61)
4	L3	0.70	(0.45)
5	L3	0.75	(0.35)
6	L3	0.40	(0.21)
7	L3	1.63	(0.62)
8	L3	1.39	(0.60)
9	L3	1.07	(0.50)
10	L3	-0.29	(1.73)
11	L3	0.92	(0.48)
12	L3	1.25	(0.63)
13	L3	0.75	(0.80)
14	L3	0.85	(0.52)
15	L3	0.08	(NaN)
16	L3	1.21	(0.68)
17	L3	0.37	(0.57)
18	L3	1.20	(0.60)
19	L3	0.16	(0.18)
20	L3	1.26	(0.62)
21	L3	1.10	(0.45)
22	L3	1.08	(0.60)
23	L3	1.08	(0.57)
24	L3	1.84	(0.59)
25	L3	1.01	(0.66)

Berdasarkan estimasi parameter diskriminasi (a) dalam model IRT 3PL, sebagian besar butir menunjukkan daya pembeda pada kategori sedang hingga tinggi. Item 1 ($a = 1,77$), item 2 ($a = 1,77$), dan item 24 ($a = 1,84$) memiliki nilai diskriminasi tinggi, yang menunjukkan kemampuan butir dalam

membedakan peserta didik berkemampuan rendah dan tinggi secara efektif. Item 7 ($a = 1,63$) dan item 3 ($a = 1,42$) juga memperlihatkan karakteristik diskriminasi yang baik. Nilai diskriminasi yang tinggi menghasilkan kurva karakteristik item (ICC) yang lebih curam dan berkontribusi positif terhadap fungsi informasi tes, sehingga meningkatkan presisi pengukuran pada rentang kemampuan tertentu.

Namun demikian, beberapa butir menunjukkan nilai diskriminasi rendah, seperti item 6 ($a = 0,40$), item 17 ($a = 0,37$), dan item 19 ($a = 0,16$), yang mengindikasikan bahwa butir kurang sensitif dalam membedakan tingkat kemampuan peserta didik. Item 15 bahkan menunjukkan nilai diskriminasi sangat rendah ($a = 0,08$), sehingga kontribusinya terhadap informasi tes sangat minimal.

Temuan yang paling krusial adalah nilai diskriminasi negatif pada item 10 ($a = -0,29$). Secara teoretis, parameter negatif menunjukkan bahwa probabilitas menjawab benar justru lebih tinggi pada peserta didik berkemampuan rendah dibandingkan berkemampuan tinggi. Untuk memastikan apakah kondisi ini disebabkan oleh multidimensionalitas, dilakukan telaah terhadap hasil analisis faktor eksploratori (EFA). Hasil EFA menunjukkan adanya satu faktor dominan dengan eigenvalue 5,702 dan tidak ditemukan faktor kedua yang stabil (eigenvalue berikutnya 1,037), serta scree plot memperlihatkan pola penurunan tajam setelah faktor pertama. Selain itu, item 10 tidak menunjukkan loading signifikan pada faktor lain dan memiliki nilai uniqueness yang sangat tinggi, yang menunjukkan bahwa varians item tersebut hampir tidak dijelaskan oleh faktor umum. Dengan demikian, tidak terdapat bukti empiris yang mendukung adanya dimensi kedua yang kuat.



Gambar 1. Scree Plot Eigenvalue pada Analisis Faktor Eksploratori (EFA)

Berdasarkan temuan tersebut, nilai diskriminasi negatif pada item 10 lebih mungkin disebabkan oleh permasalahan konstruksi butir, seperti distraktor yang tidak berfungsi optimal atau ambiguitas redaksi soal, daripada indikasi multidimensionalitas konstruk. Butir dengan karakteristik seperti ini berpotensi menurunkan kualitas fungsi informasi tes dan mengganggu kestabilan estimasi parameter, sehingga direkomendasikan untuk direvisi secara menyeluruh atau dieliminasi sebelum tahap kalibrasi skala luas.

Secara keseluruhan, meskipun mayoritas butir telah menunjukkan daya pembeda yang memadai, keberadaan beberapa butir dengan diskriminasi rendah dan satu butir dengan diskriminasi negatif memerlukan evaluasi lebih lanjut guna memastikan kualitas pengukuran instrumen tetap optimal.

3.2 Parameter Tingkat Kesulitan (b)

Parameter tingkat kesulitan (*difficulty parameter*) pada model IRT 3PL menunjukkan posisi kesulitan suatu butir soal pada skala kemampuan peserta didik. Berdasarkan hasil estimasi, nilai parameter kesulitan (b) pada instrumen ini berkisar antara $-2,67$ hingga $5,83$.

Tabel 2. Parameter 2 PL

Item	Code	Apar	(SE)	Bpar	(SE)
1	L3	1.77	(0.61)	-0.78	(0.42)
2	L3	1.77	(0.53)	-0.24	(0.35)
3	L3	1.42	(0.61)	-0.52	(0.57)
4	L3	0.70	(0.45)	-1.14	(3.10)
5	L3	0.75	(0.35)	3.55	(1.81)
6	L3	0.40	(0.21)	2.46	(1.78)
7	L3	1.63	(0.62)	-0.88	(0.51)
8	L3	1.39	(0.60)	-0.41	(0.58)
9	L3	1.07	(0.50)	-0.21	(0.87)
10	L3	-0.29	(1.73)	0.29	(17.67)
11	L3	0.92	(0.48)	1.13	(1.03)
12	L3	1.25	(0.63)	-1.35	(0.96)
13	L3	0.75	(0.80)	4.14	(4.29)
14	L3	0.85	(0.52)	5.50	(3.03)
15	L3	0.08	(NaN)	-0.31	(NaN)
16	L3	1.21	(0.68)	-2.67	(2.29)
17	L3	0.37	(0.57)	-0.90	(5.66)
18	L3	1.20	(0.60)	-0.80	(0.76)
19	L3	0.16	(0.18)	3.32	(5.26)
20	L3	1.26	(0.62)	-0.04	(0.74)
21	L3	1.10	(0.45)	0.55	(0.64)
22	L3	1.08	(0.60)	-0.50	(0.94)
23	L3	1.08	(0.57)	-0.26	(0.89)
24	L3	1.84	(0.59)	-0.62	(0.49)
25	L3	1.01	(0.66)	5.83	(1.57)

Sebagian besar butir menunjukkan nilai parameter kesulitan (b) dalam rentang moderat, yaitu sekitar -2 hingga $+2$, yang mengindikasikan bahwa butir mampu mengukur peserta didik pada rentang kemampuan rendah hingga tinggi secara proporsional. Item dengan nilai negatif seperti item 1 ($b = -0,78$), item 7 ($b = -0,88$), dan item 12 ($b = -1,35$) tergolong relatif mudah karena probabilitas menjawab benar sudah tinggi pada peserta didik dengan kemampuan di bawah rata-rata. Item 16 bahkan memiliki nilai $b = -2,67$ yang menunjukkan bahwa soal tersebut sangat mudah dan hanya memberikan informasi pada kemampuan sangat rendah, sehingga kontribusinya terhadap pengukuran pada rentang kemampuan menengah dan tinggi menjadi terbatas.

Sebaliknya, ditemukan beberapa item dengan nilai kesulitan yang sangat ekstrem, yaitu item 5 ($b = 3,55$), item 13 ($b = 4,14$), item 14 ($b = 5,50$), dan item 25 ($b = 5,83$). Dalam konteks IRT, nilai b di atas $+3$ umumnya dianggap sangat sulit dan berada di luar rentang kemampuan mayoritas responden. Item dengan nilai ekstrem seperti ini cenderung jarang dijawab benar dan menghasilkan kurva karakteristik item yang bergeser jauh ke kanan. Akibatnya, kontribusi item terhadap fungsi informasi tes menjadi sangat sempit dan hanya memberikan informasi pada kemampuan sangat tinggi yang mungkin tidak terwakili secara memadai dalam distribusi sampel.

Nilai kesulitan yang terlalu ekstrem juga dapat menjadi indikasi adanya permasalahan konstruksi butir, seperti redaksi soal yang terlalu kompleks, tuntutan kognitif yang tidak proporsional dengan indikator, atau ketidaksesuaian antara konten soal dan pengalaman belajar peserta didik. Selain itu, jika distribusi kemampuan responden tidak mencakup rentang kemampuan yang cukup tinggi, estimasi parameter kesulitan dapat menjadi tidak stabil. Oleh karena itu, item dengan nilai b di atas $+3$ direkomendasikan untuk ditinjau kembali dan dipertimbangkan untuk direvisi atau dieliminasi sebelum tahap kalibrasi skala luas.

Secara keseluruhan, meskipun sebagian besar butir telah berada pada kategori kesulitan yang memadai, keberadaan beberapa item dengan tingkat kesulitan sangat ekstrem berpotensi mengurangi efisiensi fungsi informasi tes dan perlu mendapat perhatian dalam penyempurnaan instrumen.

3.3. Parameter Faktor Tebakan (c)

Parameter guessing (c parameter) dalam model IRT 3PL menggambarkan peluang peserta didik dengan kemampuan sangat rendah untuk menjawab butir soal dengan benar melalui tebakan. Berdasarkan hasil estimasi, nilai parameter guessing pada instrumen ini berkisar antara $0,14$ hingga $0,50$.

Tabel 3. Parameter 3 PL

Item	Code	Apar	(SE)	Bpar	(SE)	Cpar	(SE)	Upar	(SE)
1	L3	1.77	(0.61)	-0.78	(0.42)	0.50	(0.14)		
2	L3	1.77	(0.53)	-0.24	(0.35)	0.14	(0.08)		
3	L3	1.42	(0.61)	-0.52	(0.57)	0.50	(0.15)		
4	L3	0.70	(0.45)	-1.14	(3.10)	0.50	(0.57)		
5	L3	0.75	(0.35)	3.55	(1.81)	0.50	(0.10)		
6	L3	0.40	(0.21)	2.46	(1.78)	0.50	(0.12)		
7	L3	1.63	(0.62)	-0.88	(0.51)	0.50	(0.16)		
8	L3	1.39	(0.60)	-0.41	(0.58)	0.50	(0.15)		
9	L3	1.07	(0.50)	-0.21	(0.87)	0.50	(0.17)		
10	L3	-0.29	(1.73)	0.29	(17.67)	0.50	(0.89)		
11	L3	0.92	(0.48)	1.13	(1.03)	0.50	(0.12)		
12	L3	1.25	(0.63)	-1.35	(0.96)	0.50	(0.28)		
13	L3	0.75	(0.80)	4.14	(4.29)	0.24	(0.09)		
14	L3	0.85	(0.52)	5.50	(3.03)	0.50	(0.09)		
15	L3	0.08	(NaN)	-0.31	(NaN)	0.50	(0.06)		
16	L3	1.21	(0.68)	-2.67	(2.29)	0.50	(0.93)		
17	L3	0.37	(0.57)	-0.90	(5.66)	0.50	(0.45)		
18	L3	1.20	(0.60)	-0.80	(0.76)	0.50	(0.18)		
19	L3	0.16	(0.18)	3.32	(5.26)	0.25	(0.11)		
20	L3	1.26	(0.62)	-0.04	(0.74)	0.50	(0.16)		
21	L3	1.10	(0.45)	0.55	(0.64)	0.50	(0.12)		
22	L3	1.08	(0.60)	-0.50	(0.94)	0.50	(0.20)		
23	L3	1.08	(0.57)	-0.26	(0.89)	0.50	(0.18)		
24	L3	1.84	(0.59)	-0.62	(0.49)	0.50	(0.16)		
25	L3	1.01	(0.66)	5.83	(1.57)	0.19	(0.06)		

Hasil estimasi parameter tebakan (c) menunjukkan bahwa sebagian item memiliki nilai yang relatif tinggi, bahkan mendekati atau mencapai 0,50. Dalam model 3PL, parameter c merepresentasikan probabilitas peserta didik dengan kemampuan sangat rendah untuk menjawab benar suatu butir. Pada soal pilihan ganda dengan empat opsi jawaban, probabilitas tebak acak secara teoretis adalah 0,25. Oleh karena itu, nilai c yang jauh melebihi 0,25 mengindikasikan adanya anomali dalam karakteristik butir.

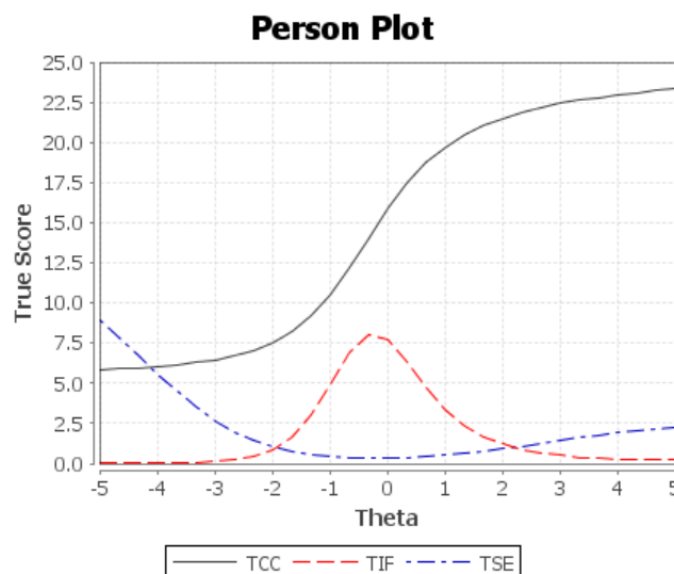
Nilai parameter c sebesar 0,50 berarti peserta didik berkemampuan sangat rendah memiliki peluang 50% untuk menjawab benar, yang secara substantif sulit dijustifikasi secara teoretis. Kondisi ini dapat disebabkan ketidakstabilan estimasi akibat ukuran sampel yang terbatas. Dalam beberapa kasus, nilai c yang tinggi juga dapat mengindikasikan bahwa model 3PL kurang sesuai dengan karakteristik data dan bahwa model yang lebih sederhana (misalnya 2PL) mungkin lebih stabil.

Dari perspektif fungsi informasi tes, item dengan nilai c yang tinggi cenderung memiliki kurva karakteristik item (ICC) yang terangkat pada bagian bawah, sehingga mengurangi kemiringan efektif kurva dan menurunkan kontribusi diskriminatif butir. Akibatnya, presisi pengukuran pada rentang kemampuan rendah menjadi kurang optimal dan dapat memengaruhi kestabilan estimasi parameter lainnya.

Dengan mempertimbangkan temuan tersebut, item dengan nilai parameter c yang sangat tinggi perlu ditinjau kembali secara substantif, khususnya pada kualitas distraktor dan kejelasan redaksi soal. Revisi atau eliminasi butir dengan nilai c ekstrem direkomendasikan sebelum tahap kalibrasi skala luas untuk memastikan kestabilan model dan peningkatan kualitas fungsi informasi tes secara keseluruhan.

3.4 Analisis Kurva Karakteristik Butir (Item Characteristic Curve/ICC)

Kurva karakteristik item (Item Characteristic Curve/ICC) digunakan untuk menggambarkan hubungan antara tingkat kemampuan peserta didik (θ) dan probabilitas menjawab benar pada setiap butir soal. Melalui grafik ini, kualitas butir dapat dianalisis berdasarkan bentuk kurva yang merepresentasikan parameter daya pembeda, tingkat kesulitan, dan peluang menebak.



Gambar 2. Grafik ICC

Analisis kurva karakteristik butir (Item Characteristic Curve/ICC) menunjukkan variasi bentuk kurva yang mencerminkan perbedaan parameter kesulitan dan diskriminasi antarbutir. Item dengan tingkat kesulitan rendah ($b < -2$) memperlihatkan kurva yang bergeser ke kiri, sehingga probabilitas menjawab benar sudah tinggi pada tingkat kemampuan rendah. Kurva pada item ini cenderung mencapai plateau lebih awal, yang berarti butir memberikan informasi terutama pada rentang kemampuan rendah dan kurang efektif membedakan peserta didik pada kemampuan menengah hingga tinggi. Sebaliknya, item dengan tingkat kesulitan sangat tinggi ($b > +3$) menunjukkan kurva yang bergeser ke kanan, dengan peningkatan probabilitas menjawab benar baru terjadi pada tingkat kemampuan sangat tinggi. Pada kondisi ini, butir hanya informatif bagi peserta didik dengan kemampuan tinggi, dan kontribusinya terhadap pengukuran menjadi terbatas apabila distribusi kemampuan responden tidak mencakup rentang tersebut.

Variasi parameter diskriminasi juga memengaruhi kemiringan kurva ICC. Item dengan nilai diskriminasi tinggi menghasilkan kurva yang lebih curam di sekitar titik kesulitan, sehingga mampu membedakan kemampuan peserta didik secara lebih tajam. Sebaliknya, item dengan diskriminasi rendah menghasilkan kurva yang lebih landai, yang menunjukkan kemampuan diferensiasi yang lemah (Hori et al., 2022).

Evaluasi lebih lanjut melalui Test Information Function (TIF) memperlihatkan bahwa puncak informasi tes berada pada rentang kemampuan sedang (sekitar θ mendekati 0). Hal ini menunjukkan bahwa instrumen memberikan presisi pengukuran tertinggi pada peserta didik dengan kemampuan menengah. Sejalan dengan itu, Standard Error of Measurement (SEM) berada pada titik terendah pada rentang tersebut dan meningkat pada kemampuan sangat rendah maupun sangat tinggi. Pola ini mengindikasikan bahwa akurasi pengukuran menurun pada kemampuan ekstrem, baik rendah maupun tinggi.

Secara keseluruhan, kombinasi analisis ICC, TIF, dan SEM menunjukkan bahwa instrumen lebih optimal dalam mengukur peserta didik pada rentang kemampuan sedang, sementara efektivitas pengukuran pada rentang kemampuan ekstrem dipengaruhi oleh keberadaan item dengan tingkat kesulitan sangat rendah, sangat tinggi, maupun nilai parameter guessing yang besar (Zein & Akhtar, 2025). Temuan ini memberikan gambaran komprehensif mengenai rentang kemampuan di mana instrumen bekerja paling efektif.

4. KESIMPULAN

Analisis menggunakan model IRT 3PL menunjukkan bahwa instrumen literasi sains memiliki karakteristik butir yang bervariasi pada parameter diskriminasi, kesulitan, dan guessing. Sebagian besar butir berada pada kategori moderat dan mampu mengukur kemampuan peserta didik secara memadai, khususnya pada rentang kemampuan sedang. Namun, ditemukan beberapa butir dengan diskriminasi rendah, satu butir dengan diskriminasi negatif, tingkat kesulitan sangat ekstrem, serta nilai guessing yang relatif tinggi, yang berpotensi mengurangi presisi pengukuran.

Dari 25 butir yang dianalisis, sebagian besar dapat dipertahankan dengan revisi, terutama pada perbaikan distraktor, sedangkan butir dengan diskriminasi negatif direkomendasikan untuk dieliminasi atau direkonstruksi. Instrumen menunjukkan potensi yang baik, tetapi belum sepenuhnya optimal untuk asesmen skala luas karena fungsi informasi tes masih terkonsentrasi pada kemampuan sedang.

Penelitian ini masih terbatas pada ukuran sampel uji coba, sehingga kalibrasi lanjutan dengan jumlah responden yang lebih besar dan distribusi kemampuan yang lebih representatif diperlukan untuk memastikan stabilitas parameter dan kelayakan penggunaan instrumen pada skala yang lebih luas.

DAFTAR PUSTAKA

- Adianto, T., & Rusli, M. A. (2021). Analysis of StudentTMs Difficulties in Solving Physics Problem: Impulse and Momentum Topics. *Unnes Science Education Journal*, 10(1), 24–33. <https://doi.org/10.15294/usej.v10i1.41517>
- Aksu, G., Güzeller, C. O., & Eser, M. T. (2019). Jmetrik: Classical test theory and item response theory data analysis software. *Journal of Measurement and Evaluation in Education and Psychology*, 10(2), 165–178. <https://doi.org/10.21031/epod.483396>
- Ariyadi, D. J. (2025). Application of Three-Parameter Logistic (3PL) Item Response Theory in Learning Management System (LMS) for Post-Test Analysis. *Journal of Informatics Development*, 3(2), 33–46. <https://doi.org/10.30741/jid.v3i2.1554>
- Gómez-Benito, J., Sireci, S., Padilla, J. L., Dolores Hidalgo, M., & Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema*, 30(1), 104–109. <https://doi.org/10.7334/psicothema2017.183>
- Gormally, C., Brickman, P., & Lut, M. (2012). Developing a test of scientific literacy skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments. *CBE Life Sciences Education*, 11(4), 364–377. <https://doi.org/10.1187/cbe.12-03-0026>
- Hooper, M. A., Tomarken, A., & Gauthier, I. (2025). Measuring visual ability in linguistically diverse populations. *Behavior Research Methods*, 57(1). <https://doi.org/10.3758/s13428-024-02579-x>
- Hori, K., Fukuhara, H., & Yamada, T. (2022). Item response theory and its applications in educational measurement Part II: Theory and practices of test equating in item response theory. In *Wiley Interdisciplinary Reviews: Computational Statistics* (Vol. 14, Number 3). John Wiley and Sons Inc. <https://doi.org/10.1002/wics.1543>
- Kalkan, Ö. K., Altun, A., & Atar, B. (2020). Role of teacher-related factors and educational resources in science literacy: An international perspective. *Studies in Educational Evaluation*, 67. <https://doi.org/10.1016/j.stueduc.2020.100935>
- Knekta, E., Rowland, A. A., Corwin, L. A., & Eddy, S. (2020). Measuring university students' interest in biology: evaluation of an instrument targeting Hidi and Renninger's individual interest. *International Journal of STEM Education*, 7(1). <https://doi.org/10.1186/s40594-020-00217-4>
- Mannan, A., Suparto, S., & Kusaeri, K. (2025). Practices and Challenges of the Validity of Exploratory Factor Analysis (EFA)-Based Assessment Instruments: A Systematic Literature Review 2020–2025. *EDUKASIA Jurnal Pendidikan Dan Pembelajaran*, 6(1), 611–626. <https://doi.org/10.62775/edukasia.v6i1.1463>
- Ocy, D. R., Sarifah, I., & Riyadi, R. (2025). EFA and CFA analysis: Development and validation of a test instrument for mathematical abstraction skills. *JRAMathEdu (Journal of Research and Advances in Mathematics Education)*. <https://doi.org/10.23917/jramathedu.v10i2.7601>

- PISA 2022 Assessment and Analytical Framework*. (2023). OECD Publishing. <https://doi.org/10.1787/dfe0bf9c-en>
- PISA 2022 Results (Volume I)*. (2023). OECD Publishing. <https://doi.org/10.1787/53f23881-en>
- Rusilowati, A., Darsono, T., Sekarningtias, F. O., & Fariyani, Q. (2024). Development of A Four-Tier E-Diagnostic Test on The Topic of Momentum to Measure and Reduce Student Misconception. *Jurnal Pendidikan Fisika Indonesia*, 20(2), 173–187. <https://doi.org/10.15294/jpfi.v20i2.1297>
- Yusmar, F., & Fadilah, R. E. (2023). Analisis Rendahnya Literasi Sains Peserta Didik Indonesia: Hasil Pisa Dan Faktor Penyebab. *Lensa (Lentera Sains): Jurnal Pendidikan IPA*, 13(1), 11–19. <https://doi.org/10.24929/lensa.v13i1.283>
- Zein, R. A., & Akhtar, H. (2025). Getting started with the graded response model: An introduction and tutorial in R. *International Journal of Psychology*, 60(1). <https://doi.org/10.1002/ijop.13265>