Jurnal Pendidikan Biologi
Universitas Negeri Malang, Jawa Timur, Indonesia
p-ISSN 2085-6873 | e-ISSN 2540-9271
Edisi Agustus 2025, Volume 16, Nomor 2, pp 155-167

ANALISIS VALIDITAS BUTIR DAN KETEPATAN TARGET UKUR INSTRUMEN BIOLOGI KELAS XII: APLIKASI PEMODELAN RASCH

Endah Pratiwi, Bowo Sugiharto*
Program Studi Pendidikan Biologi, Universitas Sebelas Maret, Jl. Ir. Sutami No.36, Jebres, Kec. Jebres, Kota Surakarta, Jawa Tengah 57126
* corresponding author | email : bowo@fkip.uns.ac.id

Received: 8 Juli 2025 *Accepted: 28 Agustus 2025* *Published: 31 Agustus 2025*

ABSTRAK

Penelitian ini bertujuan menganalisis kualitas instrumen tes bioteknologi terdiri dari 20 butir soal menggunakan Model Rasch. Analisis item menunjukkan terdapat 3 butir misfit (S7, S14, S19) yang melampaui batas Outfit MNSQ sehingga tidak sesuai dengan prediksi model. Distribusi kesukaran soal berada pada rentang –2,94 hingga +5,11 logit, menunjukkan variasi kompleksitas yang luas namun cenderung terpolarisasi pada kategori sangat mudah dan sangat sukar. Pada tingkat responden, ditemukan 11 siswa (22%) dengan pola respons misfit yang mengindikasikan kemungkinan guessing atau ketidakkonsistenan respons. Meskipun reliabilitas item dan person berada pada kategori baik, pola ketidakcocokan ini menandakan perlunya perbaikan instrumen. Rekomendasi utama penelitian mencakup revisi butir misfit, penambahan soal pada rentang kesukaran moderat untuk menutup gap pada kemampuan ekstrem, serta penyusunan ulang blueprint agar distribusi tingkat kesukaran lebih proporsional. Hasil penelitian menegaskan pentingnya kalibrasi instrumen berbasis Rasch untuk menghasilkan evaluasi pembelajaran yang lebih akurat dan objektif.

doi http://dx.doi.org/10.17977/um052v16i2p155-167

Kata Kunci : *Model Rasch; analisis butir; butir misfit; tingkat kesukaran logit; kecocokan responden; kalibrasi tes; asesmen bioteknologi; validitas pengukuran; evaluasi pembelajaran*

This study aimed to analyze the quality of a 20-item biotechnology test using the Rasch Model. The item analysis identified three misfitting items (S7, S14, S19) whose Outfit MNSQ values exceeded the model's acceptable range. The difficulty levels ranged from –2.94 to +5.11 logits, indicating a wide spread of item complexity but with a tendency toward extreme categories, dominated by very easy and very difficult items. At the respondent level, 11 students (22%) showed misfitting response patterns, suggesting potential guessing or inconsistent answering behavior. Although item and person reliability were classified as good, the presence of misfit patterns indicates the need for instrument refinement. Key recommendations include revising the misfitting items, adding more items within the moderate difficulty range to fill gaps at the extremes, and restructuring the test blueprint to achieve a more balanced distribution of difficulty. The findings highlight the importance of Rasch-based calibration to produce more accurate and objective educational assessments.

Keywords: *Rasch Model; item analysis; misfit items; logit difficulty; person fit; test calibration; biotechnology assessment; measurement validity; educational evaluation*

Era pendidikan modern yang menuntut penguasaan kompetensi abad ke-21, evaluasi pembelajaran memegang peranan krusial. Evaluasi tidak sekadar bertujuan untuk memberikan label nilai kuantitatif kepada siswa, melainkan berfungsi sebagai alat diagnostik untuk memastikan ketercapaian tujuan pembelajaran secara menyeluruh (Utami et al., 2025). Kualitas evaluasi sangat bergantung pada kualitas instrumen yang digunakan. Instrumen penilaian yang valid dan reliabel merupakan prasyarat mutlak untuk menghasilkan data yang akurat sebagai dasar pengambilan keputusan pendidikan, perbaikan strategi pengajaran, serta pemetaan kemampuan peserta didik yang

objektif (Syafi'i et al, 2025). Tanpa instrumen yang terstandarisasi, hasil pengukuran berisiko bias dan gagal menilai kompetensi siswa yang sesungguhnya. Pendekatan ini juga menjamin bahwa instrumen pengukuran hanya mengukur satu atribut kemampuan dominan siswa, meminimalkan gangguan dari variabel lain yang tidak relevan(Fitraynsyah & Hilmiyati, 2024)

Namun, terdapat kesenjangan (gap) yang nyata dalam praktik evaluasi pembelajaran di sekolah. Pada mata pelajaran Biologi tingkat SMA, sebagian besar soal ulangan harian disusun secara internal oleh guru dan langsung digunakan tanpa proses uji validitas maupun analisis empiris. Akibatnya, sering ditemukan ketidaksesuaian (mismatch) antara tingkat kesukaran soal dan kemampuan siswa: siswa berkemampuan tinggi memperoleh skor rendah, sementara siswa berkemampuan rendah justru lulus karena dominasi soal sangat mudah. Fenomena ini mengindikasikan adanya defisiensi psikometrik seperti indikator tidak selaras, redaksi ambigu, dan pola respons tidak konsisten (Rivaldi, 2025). Selain itu, kualitas butir soal sangat penting dalam menentukan sejauh mana tes tersebut mampu memberikan gambaran akurat tentang kemampuan siswa (Farida dan Anna, 2021).

Temuan penelitian terbaru turut memperkuat permasalahan tersebut. Studi tahun 2022–2024 menunjukkan bahwa instrumen buatan guru di sekolah sering mengalami *ceiling effect* dan *floor effect*, yaitu skor siswa menumpuk di nilai sangat tinggi atau sangat rendah sehingga kemampuan siswa tidak dapat dibedakan secara (Ahmad et al., 2025). Selain itu, banyak instrumen sekolah mengalami *targeting problem*, yaitu ketidaksesuaian distribusi tingkat kesulitan soal dengan kemampuan rata-rata siswa sehingga mengurangi daya diskriminasi instrumen (Irwanto, 2025). Kualitas instrumen yang tidak optimal ini menyebabkan guru gagal memperoleh profil kemampuan siswa yang sebenarnya.

Untuk mengatasi permasalahan pengukuran tersebut, diperlukan pendekatan analisis yang mampu mendeteksi karakteristik butir soal secara mendalam, melampaui batasan Teori Tes Klasik (*Classical Test Theory*). Pendekatan yang dinilai paling komprehensif adalah *Item Response Theory* (IRT), khususnya Pemodelan Rasch (*Rasch Model*). Berbeda dengan teori klasik yang sangat bergantung pada sampel, Model Rasch memiliki sifat *sample-free* dan menawarkan kalibrasi yang setara antara tingkat kesulitan soal dan kemampuan siswa dalam satu garis linear atau skala logit (Boone et al., 2014). Keunggulan utama Model Rasch terletak pada kemampuannya menyajikan *Wright Map*, sebuah peta visual yang memudahkan pendidik mendeteksi *misfit* (ketidakcocokan) respon, bias butir soal (DIF), serta memisahkan butir yang terlalu mudah atau terlalu sulit yang tidak memberikan informasi signifikan terhadap kemampuan (Novriyanti & Arthur, 2024).

Transformasi data ke dalam skala logit ini menjadi aspek fundamental karena menjamin properti objektivitas pengukuran yang tidak dimiliki oleh pendekatan klasik. Dalam Model Rasch, probabilitas keberhasilan siswa mengerjakan soal hanya ditentukan oleh selisih antara kemampuan siswa dan kesulitan butir, sehingga data ordinal dari skor mentah dikonversi menjadi data interval yang memiliki satuan ukur setara (Dewi et al., 2018). Hal ini memungkinkan pendidik untuk memastikan bahwa instrumen tes memiliki invariansi pengukuran, artinya alat ukur tersebut tetap berfungsi konsisten dan stabil meskipun digunakan pada kelompok siswa yang berbeda karakteristiknya

Meskipun Model Rasch menawarkan presisi diagnostik yang superior dalam memvalidasi instrumen tes buatan guru, implementasinya dalam evaluasi pendidikan di tingkat sekolah masih belum optimal (Ibrahim, 2025). Padahal, penerapan analisis probabilitas ini krusial untuk mengidentifikasi defisiensi psikometrik pada butir soal yang bersifat laten, guna menjamin bahwa skor asesmen merefleksikan kemampuan kognitif (*latent trait*) siswa secara valid, serta meminimalisir bias pengukuran akibat faktor tebakan (*guessing*) maupun ambiguitas konstruksi soal.

Berdasarkan uraian permasalahan di atas, penelitian ini bertujuan untuk menganalisis kualitas butir soal mata pelajaran bioteknologi siswa kelas XII**.** Secara spesifik, penelitian ini bertujuan untuk: (1) menguji validitas dan reliabilitas instrumen; (2) Menganalisis tingkat kesukaran butir soal serta (3) mendeteksi adanya butir soal yang bias atau tidak konsisten (*misfit*). Hasil analisis ini diharapkan dapat memberikan rekomendasi konkret bagi perbaikan bank soal di sekolah tersebut.

METODE

Penelitian ini merupakan penelitian deskriptif kuantitatif yang bertujuan untuk menganalisis karakteristik psikometrik instrumen tes evolusi dan memetakan kemampuan siswa menggunakan pendekatan Model Rasch. Subjek penelitian terdiri atas 50 siswa kelas XII dari salah satu SMA Negeri di Kabupaten Karanganyar yang telah menyelesaikan materi evolusi pada semester genap. Pengambilan sampel dilakukan secara *purposive sampling* dengan mempertimbangkan keterjangkauan lokasi, kesediaan sekolah untuk berpartisipasi, serta kesesuaian kurikulum yang digunakan. Instrumen penelitian berupa tes pilihan ganda sebanyak 20 butir soal yang disusun oleh guru biologi berdasarkan Kompetensi Dasar KD 3.11, yaitu "Menganalisis mekanisme evolusi serta bukti-bukti pendukung terjadinya evolusi".

Penyusunan soal mengacu pada delapan indikator pembelajaran yang mencakup kemampuan menjelaskan konsep dasar evolusi, mengidentifikasi bukti-bukti evolusi, membedakan mekanisme seleksi alam, menjelaskan peran mutasi, aliran gen, dan rekombinasi genetik dalam evolusi, menganalisis hubungan filogenetik, serta menerapkan konsep Hardy–Weinberg dalam menyelesaikan persoalan populasi. Butir soal dikembangkan dengan rentang level kognitif C1–C4, terdiri atas 4 soal C1 (mengingat), 6 soal C2 (memahami), 6 soal C3 (menerapkan), dan 4 soal C4 (menganalisis). Contoh butir pada level C3 misalnya, "Dalam suatu populasi, perubahan frekuensi alel dapat terjadi karena peristiwa perpindahan individu dari satu populasi ke populasi lain. Selanjutnya yang ditanyakan adalah mekanisme evolusi yang digunakan. Data respon siswa dikodekan secara dikotomus (benar = 1; salah = 0) dan dianalisis menggunakan perangkat lunak Winsteps berbasis Model Rasch.

Analisis Rasch difokuskan pada pemeriksaan kesesuaian butir (item fit) dan kesesuaian responden (person fit), kalibrasi tingkat kesulitan butir dan kemampuan siswa dalam skala logit, serta visualisasi keduanya pada *Wright Map.* Kriteria kelayakan butir mengikuti standar Rasch, yaitu nilai Outfit MNSQ berada pada rentang 0,5–1,5, nilai ZSTD –2,0 hingga +2,0, serta *Point Measure Correlation* bernilai positif dengan batas ideal > 0,30. Kriteria yang sama diterapkan untuk mendeteksi person misfit, sehingga responden dengan nilai Outfit MNSQ atau ZSTD di luar batas toleransi dikategorikan memiliki pola jawaban tidak konsisten. Kualitas instrumen dievaluasi menggunakan reliabilitas item, reliabilitas person, dan Cronbach's Alpha (KR-20) untuk menilai konsistensi internal. Selain itu, analisis juga mempertimbangkan adanya fenomena *ceiling effect* (butir terlalu mudah), *floor effect* (butir terlalu sulit), serta kesesuaian atau ketidaksejajaran (*targeting*) antara distribusi kemampuan siswa dan tingkat kesulitan butir. Penelitian ini mengasumsikan bahwa seluruh responden mengerjakan tes secara jujur dan memperhatikan kemungkinan keterbatasan pada ukuran sampel yang digunakan.

HASIL DAN PEMBAHASAN

Validitas Item (Soal)

Analisis validitas butir berdasarkan model Rasch dilakukan dengan memeriksa parameter *infit* dan *outfit mean square* (MNSQ) sebagai indikator utama kesesuaian butir terhadap model pengukuran dapat dilihat pada Gambar 1. Rentang nilai MNSQ yang dinyatakan produktif untuk pengukuran terletak antara 0.5 hingga 1.5, yang menunjukkan bahwa butir bekerja secara konsisten dalam mengukur konstruk yang dimaksud (Akram et al., 2015). Berdasarkan output, mayoritas butir dalam instrumen memiliki nilai *infit* dan *outfit MNSQ* yang berada dalam batas tersebut, sehingga mengindikasikan tingkat kesesuaian yang memadai dengan model Rasch. Namun demikian, beberapa item seperti S7, S14, dan S19 menunjukkan nilai *outfit MNSQ* yang melampaui batas atas, menandakan adanya *underfit* atau respons yang tidak konsisten dengan prediksi model. Kondisi ini menunjukkan bahwa butir-butir tersebut berpotensi mengukur aspek yang berbeda atau mengandung pola respons yang tidak stabil.

Validitas konstruk juga ditinjau melalui parameter *point measure correlation* (PTMEA CORR), yang merefleksikan sejauh mana respons terhadap suatu item berkorelasi secara positif dengan kemampuan peserta secara keseluruhan. Kisaran nilai PTMEA CORR yang direkomendasikan adalah 0.20–0.80, yang menunjukkan bahwa butir memberikan kontribusi yang konsisten terhadap

pengukuran (Bangi, 2022). Seluruh butir pada instrumen menunjukkan nilai PTMEA CORR positif dan berada dalam rentang yang dapat diterima, sehingga menegaskan bahwa tidak terdapat indikasi butir terbalik (*miskeyed item*) maupun item yang mengukur konstruk yang berbeda dari konstruk utama. Dengan demikian, dari perspektif hubungan antara item dan kemampuan responden, instrumen ini menunjukkan koherensi internal yang kuat dan mendukung validitas konstruknya.

Selain itu, karakteristik tingkat kesukaran butir yang tercermin dalam nilai *measure* memperlihatkan rentang yang cukup luas, yaitu dari –2.94 hingga +5.11, menandakan bahwa instrumen mencakup variasi kompleksitas yang memadai, mulai dari butir yang relatif mudah hingga sangat sulit. Distribusi tingkat kesukaran yang demikian adalah indikator penting bahwa instrumen mampu membedakan kemampuan peserta secara efektif di berbagai level. Rata-rata nilai *infit* dan *outfit MNSQ* yang berada dekat dengan angka ideal (≈1.00) semakin menguatkan bahwa instrumen memenuhi asumsi model Rasch secara keseluruhan. Dengan mempertimbangkan kesesuaian model, konsistensi korelasi butir, serta distribusi tingkat kesukaran, dapat disimpulkan bahwa instrumen ini memiliki validitas konstruksi yang kuat, meskipun beberapa butir dengan pola *misfit* yang menonjol tetap direkomendasikan untuk direvisi guna meningkatkan kualitas pengukuran.

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
7	3	50	5.11	.62	1.26	.66	2.31	1.26	A -.05	.24	94.0	93.9	S7
19	14	50	3.01	.35	1.47	2.69	1.75	1.78	B .08	.42	54.0	75.0	S19
14	39	50	.09	.39	1.31	1.38	.91	-.06	C .34	.47	66.0	81.6	S14
1	44	50	-.83	.48	1.21	.77	.89	.07	D .29	.39	82.0	88.0	S1
11	27	50	1.58	.33	1.18	1.25	1.06	.33	E .41	.50	56.0	72.5	S11
15	45	50	-1.07	.51	1.12	.45	.53	-.43	F .36	.36	90.0	89.9	S15
18	48	50	-2.18	.75	1.08	.33	.66	.11	G .20	.24	96.0	95.9	S18
3	49	50	-2.94	1.02	1.07	.38	.97	.45	H .09	.17	98.0	98.0	S3
13	39	50	.09	.39	1.00	.07	.98	.10	g .46	.47	86.0	81.6	S13
2	46	50	-1.36	.56	.99	.09	.39	-.58	f .40	.33	92.0	91.9	S2
4	30	50	1.25	.34	.91	-.55	.80	-.83	e .58	.51	74.0	74.7	S4
17	44	50	-.83	.48	.66	-1.20	.34	-1.04	d .60	.39	94.0	88.0	S17
20	47	50	-1.71	.63	.66	-.68	.19	-.91	c .51	.29	94.0	93.9	S20
16	46	50	-1.36	.56	.61	-1.05	.21	-1.02	b .57	.33	92.0	91.9	S16
12	31	50	1.14	.34	.53	-3.38	.44	-2.81	a .81	.51	88.0	75.4	S12
MEAN	37.6	50.0	-.46	.84	1.01	.08	.83	-.24			83.7	86.2	
P.SD	15.2	.0	3.05	.59	.27	1.34	.56	1.04			14.0	8.3	

Gambar 1. Misfit Order

Validitas Responden (*Person Fit)*

Evaluasi kesesuaian responden (*person fit*) merupakan komponen krusial dalam analisis psychometric berbasis Model Rasch untuk memverifikasi sejauh mana pola respons individu selaras dengan prediksi model probabilistik. Konsistensi respons mengindikasikan bahwa data yang terekam secara valid merefleksikan atribut latent yang diukur (Akram et al., 2015). Sebaliknya, deviasi signifikan dari ekspektasi model dapat menjadi indikator adanya fenomena respons yang tidak lazim, seperti *random guessing*, *carelessness*, atau *fatigue*, yang berpotensi mendistorsi estimasi kemampuan (Riskawati et al., 2025). Dalam penelitian ini, kriteria untuk menilai kesesuaian responden didasarkan pada rentang statistik *fit* yang direkomendasikan yaitu, *Outfit Mean Square* (MNSQ) dalam rentang 0,5 hingga 1,5 dan *Outfit Z-Standard* (ZSTD) antara -2,0 hingga +2,0, sesuai praktik terbaik dalam pengukuran Rasch (Sumintono, B., & Widhiarso, 2015).

Analisis *Person Statistics* dari 50 siswa kelas XII yang disajikan pada Gambar 2. Bahwa mayoritas subjek (39 dari 50 siswa, atau 78%) menunjukkan pola respons yang koheren dengan Model Rasch. Kelompok responden ini memperlihatkan nilai *Outfit MNSQ* dan *ZSTD* yang berada dalam batas

keberterimaan, misalnya siswa dengan *Entry Number* 37 (Person 56L) yang memiliki MNSQ 0,86 dan ZSTD -0,81. Pola respons yang konsisten ini menegaskan bahwa siswa-siswa tersebut merespons tes secara serius dan jawaban mereka valid merepresentasikan tingkat kemampuan biologi yang sebenarnya. Namun, sejumlah 11 siswa (22%) teridentifikasi sebagai *misfit*, menunjukkan inkonsistensi yang perlu dicermati. Beberapa di antaranya, seperti siswa dengan *Entry Number* 15 (Person 30L, 42P, 58P), mencatat *Outfit MNSQ* yang sangat tinggi (mencapai 4,03) dan *Outfit ZSTD* signifikan (>+2,08). Nilai MNSQ yang melebihi batas 1,5 mengindikasikan adanya *underfit*, di mana respons individu terlalu acak atau tidak terprediksi, seringkali dikaitkan dengan *guessing* pada butir-butir sulit atau kesalahan ceroboh pada butir mudah (Ryan et al., 2022). Sementara itu, tidak ditemukan responden yang secara signifikan *overfit* (MNSQ < 0,5), yang menunjukkan bahwa tidak ada pola respons yang terlalu deterministik dan kurang memberikan informasi baru. Secara agregat, meskipun terdapat proporsi kecil responden yang menyimpang, dominasi *person fit* pada sebagian besar siswa memberikan dukungan kuat terhadap validitas data respons dan inferensi mengenai kemampuan siswa dalam materi biologi.

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%	Person
37	18	20	4.48	1.35	.30	-.86	.06	-.81	.74	.68	100.0	95.3	47L
46	18	20	4.48	1.35	.30	-.86	.06	-.81	.74	.68	100.0	95.3	56L
14	17	20	3.07	1.04	.38	-.94	.13	-.58	.78	.71	93.3	91.7	14P
15	17	20	3.07	1.04	.38	-.94	.13	-.58	.78	.71	93.3	91.7	15P
24	17	20	3.07	1.04	.38	-.94	.13	-.58	.78	.71	93.3	91.7	24P
32	17	20	3.07	1.04	.38	-.94	.13	-.58	.78	.71	93.3	91.7	32P
34	17	20	3.07	1.04	.38	-.94	.13	-.58	.78	.71	93.3	91.7	34P
36	17	20	3.07	1.04	.38	-.94	.13	-.58	.78	.71	93.3	91.7	36P
39	17	20	3.07	1.04	.38	-.94	.13	-.58	.78	.71	93.3	91.7	49P
40	17	20	3.07	1.04	.38	-.94	.13	-.58	.78	.71	93.3	91.7	40L
41	17	20	3.07	1.04	.38	-.94	.13	-.58	.78	.71	93.3	91.7	41P
45	17	20	3.07	1.04	.38	-.94	.13	-.58	.78	.71	93.3	91.7	45P
47	17	20	3.07	1.04	.38	-.94	.13	-.58	.78	.71	93.3	91.7	57L
1	16	20	2.16	.87	.53	-.91	.23	-.36	.78	.71	93.3	88.6	01P
2	16	20	2.16	.87	.53	-.91	.23	-.36	.78	.71	93.3	88.6	02P
7	16	20	2.16	.87	.53	-.91	.23	-.36	.78	.71	93.3	88.6	07P
10	16	20	2.16	.87	.53	-.91	.23	-.36	.78	.71	93.3	88.6	10P
12	16	20	2.16	.87	1.54	1.05	.94	.42	.65	.71	80.0	88.6	12L
13	16	20	2.16	.87	.53	-.91	.23	-.36	.78	.71	93.3	88.6	13P
16	16	20	2.16	.87	1.54	1.05	.94	.42	.65	.71	80.0	88.6	16L
17	16	20	2.16	.87	.53	-.91	.23	-.36	.78	.71	93.3	88.6	17L
18	16	20	2.16	.87	.53	-.91	.23	-.36	.78	.71	93.3	88.6	18P
20	16	20	2.16	.87	.53	-.91	.23	-.36	.78	.71	93.3	88.6	20P
23	16	20	2.16	.87	1.54	1.05	.94	.42	.65	.71	80.0	88.6	23L
25	16	20	2.16	.87	.53	-.91	.23	-.36	.78	.71	93.3	88.6	25P
27	16	20	2.16	.87	.53	-.91	.23	-.36	.78	.71	93.3	88.6	27L
29	16	20	2.16	.87	.53	-.91	.23	-.36	.78	.71	93.3	88.6	39P
33	16	20	2.16	.87	.53	-.91	.23	-.36	.78	.71	93.3	88.6	33P
35	16	20	2.16	.87	1.54	1.05	.94	.42	.65	.71	80.0	88.6	35L
8	15	20	1.49	.78	1.41	.98	.88	.22	.66	.71	66.7	84.2	08L
19	15	20	1.49	.78	1.41	.98	.88	.22	.66	.71	66.7	84.2	29L
21	15	20	1.49	.78	.82	-.30	.49	-.26	.74	.71	93.3	84.2	21P
30	15	20	1.49	.78	2.49	2.58	4.03	2.06	.44	.71	53.3	84.2	30L
42	15	20	1.49	.78	2.49	2.58	4.03	2.06	.44	.71	53.3	84.2	42P
48	15	20	1.49	.78	2.49	2.58	4.03	2.06	.44	.71	53.3	84.2	58P
9	14	20	.92	.73	.97	.04	3.45	2.02	.66	.70	86.7	82.2	09P
11	14	20	.92	.73	1.75	1.69	2.69	1.61	.54	.70	73.3	82.2	11P
50	14	20	.92	.73	1.02	.19	.91	.19	.70	.70	86.7	82.2	50P
3	13	20	.43	.69	1.01	.14	.79	.05	.70	.69	80.0	80.6	03P
5	13	20	.43	.69	1.01	.14	.79	.05	.70	.69	80.0	80.6	05P
6	13	20	.43	.69	1.01	.14	.79	.05	.70	.69	80.0	80.6	06P
31	13	20	.43	.69	1.01	.14	.79	.05	.70	.69	80.0	80.6	31P
22	12	20	-.03	.66	2.07	2.55	2.99	1.65	.46	.68	46.7	78.5	22P
26	12	20	-.03	.66	1.21	.71	1.05	.39	.65	.68	73.3	78.5	26P
38	12	20	-.03	.66	1.13	.47	.91	.24	.67	.68	73.3	78.5	48L
49	12	20	-.03	.66	1.13	.47	.91	.24	.67	.68	73.3	78.5	59P
43	11	20	-.45	.65	.81	-.57	.60	.00	.71	.67	80.0	77.3	43P
4	10	20	-.87	.65	1.01	.12	.74	.17	.67	.66	73.3	75.4	04P
28	10	20	-.87	.65	1.01	.12	.74	.17	.67	.66	73.3	75.4	38L
44	10	20	-.87	.65	1.01	.12	.74	.17	.67	.66	73.3	75.4	44L
MEAN	15.0	20.0	1.75	.86	.91	-.06	.83	.06			83.7	86.2	
P.SD	2.1	.0	1.30	.17	.59	1.08	1.07	.77			12.9	5.4	

Gambar 2. PTMEA CORR Soal Tes

Wright Map

Peta Wright dalam analisis ini menggambarkan hubungan antara kemampuan peserta dan kesulitan item, yang memperlihatkan distribusi keduanya. Visualisasi ini sangat penting karena memungkinkan peneliti untuk mengevaluasi apakah item yang disusun mampu mencakup rentang kemampuan peserta secara proporsional. Pada sumbu vertikal peta, bagian kiri biasanya menunjukkan distribusi kemampuan peserta (person), sedangkan bagian kanan menunjukkan distribusi tingkat kesulitan soal (item).

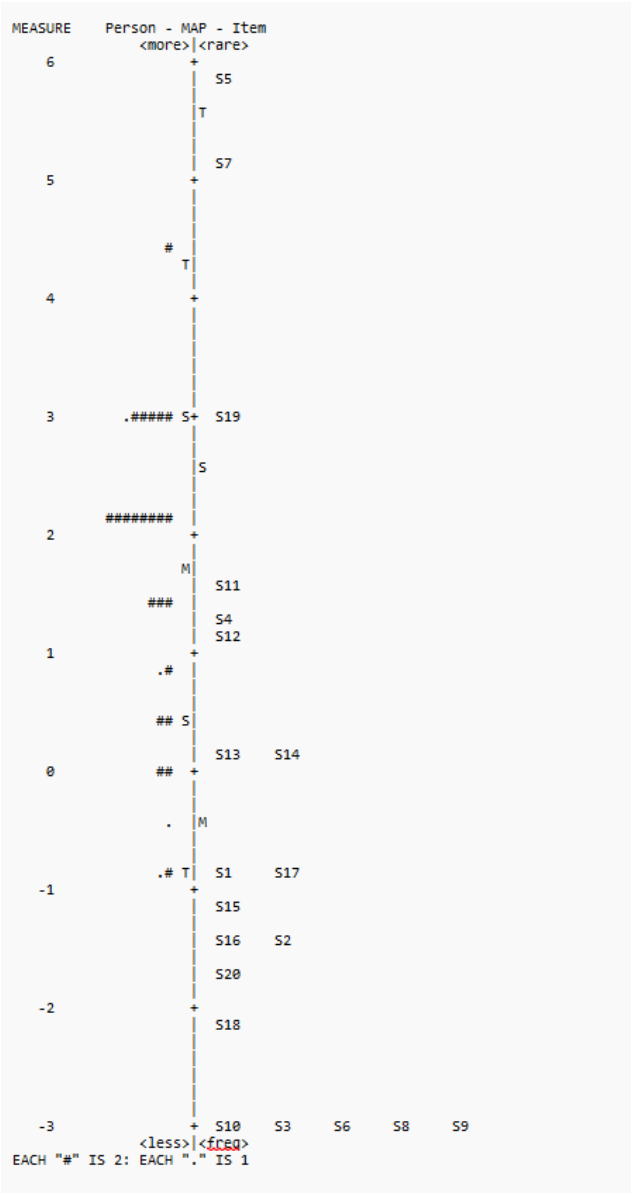
Dalam interpretasinya, nilai logit positif pada skala person menunjukkan bahwa responden memiliki kemampuan di atas rata-rata, sementara logit negatif mencerminkan kemampuan di bawah rata-rata. Sebaliknya, nilai logit positif pada skala item menunjukkan bahwa soal tersebut lebih sulit,

dan nilai negatif berarti soal tersebut lebih mudah (Zhao et al., 2023). Semakin tinggi posisi suatu titik pada skala logit, maka semakin besar pula kemampuan individu atau semakin tinggi tingkat kesulitan butir yang dimaksud (Tavakol, M., & Dennick, 2010).

Person–Item Map memberikan representasi visual mengenai hubungan antara kemampuan responden dan tingkat kesukaran item dalam satu kontinum logit, sehingga memungkinkan evaluasi kesesuaian instrumen dalam kerangka pemodelan Rasch. Distribusi kemampuan responden pada peta tersebut terlihat memusat pada rentang logit 2 hingga –1, dengan kepadatan tertinggi di sekitar logit 2–1, yang menunjukkan bahwa sebagian besar peserta berada pada kategori kemampuan menengah hingga relatif tinggi. Sementara itu, tingkat kesukaran item tersebar cukup luas, mulai dari logit +6 sebagai kategori tersulit hingga –3 sebagai kategori termudah. Item seperti S5 dan S7 menempati posisi logit tertinggi sehingga tergolong sangat sulit, sedangkan item S10, S3, S6, S8, dan S9 berada pada logit terendah dan termasuk item yang paling mudah. Penyebaran logit yang luas ini mengindikasikan bahwa instrumen mencakup variasi tingkat kesukaran yang memadai untuk membedakan kemampuan responden secara efektif (Noperi et al., 2024).

Kesesuaian antara kemampuan responden dan tingkat kesukaran butir merupakan indikator penting kualitas instrumen. Berdasarkan peta tersebut, sebagian besar item berada dalam rentang kemampuan mayoritas peserta sehingga instrumen dianggap cukup efektif dalam mengukur konstruk yang ditargetkan. Namun demikian, terdapat celah (*gap*) pada rentang kemampuan ekstrem. Pada kemampuan sangat tinggi (logit 3–6), hanya sedikit item seperti S5 dan S7 yang mewakili rentang tersebut, sehingga instrumen kurang mampu menantang peserta berkemampuan tinggi. Sebaliknya, pada rentang kemampuan sangat rendah, hanya sedikit item yang berada pada logit bawah seperti S10. Ketidakseimbangan ini menunjukkan perlunya penambahan atau penyesuaian butir untuk memperluas cakupan pengukuran, sesuai rekomendasi penelitian kontemporer tentang pengembangan instrumen berbasis Rasch

Secara keseluruhan, Person–Item Map menunjukkan bahwa instrumen memiliki kualitas pemetaan yang baik, dengan sebagian besar butir berada dalam jangkauan kemampuan responden. Meskipun demikian, kekurangan pada area kemampuan ekstrem menunjukkan adanya ruang untuk perbaikan dalam pengembangan instrumen berikutnya.



Gambar3. Peta Sebaran Individu Item

Untuk mempermudah proses interpretasi tingkat kesukaran butir dalam model Rasch, rentang nilai logit perlu diklasifikasikan ke dalam kategori tertentu. Pengelompokan ini memungkinkan peneliti menilai apakah suatu butir masuk dalam kategori sangat sukar, sukar, sedang, mudah, atau sangat mudah berdasarkan posisi nilai logitnya pada skala pengukuran. Selain itu, kategori khusus juga diberikan untuk butir yang berada di luar rentang logit yang wajar (outliers), karena butir tersebut berpotensi mengganggu keakuratan model. Adapun klasifikasi lengkap rentang logit dan kategori kesukaran disajikan pada Tabel 1.(Sumintono, B., & Widhiarso, 2015).

Tabel 1. Kriteria Tingkat Kesukaran Butir Soal

Rentang Nilai Logit (Measure)	Kategori Kesukaran
>+1.0 logit	Sangat Sukar
0.0<b≤+1.0 logit	Sukar
−1.0<b≤0.0 logit	Mudah
<−1.0 logit	Sangat Mudah

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	0	50	7.56	1.83	MAXIMUM MEASURE				.00	.00	100.0	100.0	S5
7	3	50	5.11	.62	1.26	.66	2.31	1.26	-.05	.24	94.0	93.9	S7
19	14	50	3.01	.35	1.47	2.69	1.75	1.78	.08	.42	54.0	75.0	S19
11	27	50	1.58	.33	1.18	1.25	1.06	.33	.41	.50	56.0	72.5	S11
4	30	50	1.25	.34	.91	-.55	.80	-.83	.58	.51	74.0	74.7	S4
12	31	50	1.14	.34	.53	-3.38	.44	-2.81	.81	.51	88.0	75.4	S12
13	39	50	.09	.39	1.00	.07	.98	.10	.46	.47	86.0	81.6	S13
14	39	50	.09	.39	1.31	1.38	.91	-.06	.34	.47	66.0	81.6	S14
1	44	50	-.83	.48	1.21	.77	.89	.07	.29	.39	82.0	88.0	S1
17	44	50	-.83	.48	.66	-1.20	.34	-1.04	.60	.39	94.0	88.0	S17
15	45	50	-1.07	.51	1.12	.45	.53	-.43	.36	.36	90.0	89.9	S15
2	46	50	-1.36	.56	.99	.09	.39	-.58	.40	.33	92.0	91.9	S2
16	46	50	-1.36	.56	.61	-1.05	.21	-1.02	.57	.33	92.0	91.9	S16
20	47	50	-1.71	.63	.66	-.68	.19	-.91	.51	.29	94.0	93.9	S20
18	48	50	-2.18	.75	1.08	.33	.66	.11	.20	.24	96.0	95.9	S18
3	49	50	-2.94	1.02	1.07	.38	.97	.45	.09	.17	98.0	98.0	S3
6	50	50	-4.18	1.83	MINIMUM MEASURE				.00	.00	100.0	100.0	S6
8	50	50	-4.18	1.83	MINIMUM MEASURE				.00	.00	100.0	100.0	S8
9	50	50	-4.18	1.83	MINIMUM MEASURE				.00	.00	100.0	100.0	S9
10	50	50	-4.18	1.83	MINIMUM MEASURE				.00	.00	100.0	100.0	S10
MEAN	37.6	50.0	-.46	.84	1.01	.08	.83	-.24			83.7	86.2	
P.SD	15.2	.0	3.05	.59	.27	1.34	.56	1.04			14.0	8.3	

Gambar 3. Output Item Measure

Berdasarkan output *item measure*, diperoleh nilai *measure* untuk masing-masing butir soal. Nilai-nilai tersebut kemudian diklasifikasikan ke dalam lima kategori sesuai dengan kriteria tingkat kesukaran yang tercantum pada Tabel 2. Dari proses tersebut, diperoleh hasil analisis mengenai tingkat kesukaran setiap butir soal.

Tabel 2. Tingkat Kesukaran Butir Soal

Kategori	No. Item (Butir Soal)	Jumlah	Persentase
Sangat Sukar	4, 5, 7, 11, 12, 19	6	30%
Sukar	13, 14	2	10%
Mudah	1, 17	2	10%
Sangat Mudah	2, 3, 6, 8, 9, 10, 15, 16, 18, 20	10	50%
Total		20	100%

Berdasarkan hasil analisis yang disajikan pada Tabel 2, distribusi tingkat kesukaran butir soal menunjukkan variasi yang cukup signifikan, namun cenderung didominasi oleh soal-soal dengan kategori ekstrem (sangat mudah dan sangat sukar). Data menunjukkan bahwa persentase terbesar terdapat pada kategori sangat mudah, yaitu sebanyak 50% dari total instrumen atau setara dengan 10 butir soal (nomor 2, 3, 6, 8, 9, 10, 15, 16, 18, dan 20). Tingginya proporsi pada kategori ini mengindikasikan bahwa separuh dari instrumen tes dapat dijawab dengan benar oleh mayoritas siswa, bahkan oleh siswa dengan kemampuan rendah sekalipun. Kondisi ini menunjukkan bahwa sejumlah butir soal tidak mampu memberikan informasi yang optimal dalam membedakan kemampuan peserta tes, karena soal yang terlalu mudah cenderung memiliki daya diskriminasi yang rendah (Simbolon, 2024).

Sebaliknya, proporsi terbesar kedua terdapat pada kategori Sangat Sukar dengan persentase sebesar 30% (6 butir soal), yang meliputi butir nomor 4, 5, 7, 11, 12, dan 19. Hal ini menunjukkan

adanya kesenjangan yang cukup lebar dalam instrumen, di mana soal melompat dari sangat mudah ke sangat sukar tanpa banyak gradasi di tengah. Kategori Sukar dan Mudah masing-masing hanya diwakili oleh 10% (2 butir soal), yaitu nomor 13 dan 14 untuk kategori sukar, serta nomor 1 dan 17 untuk kategori mudah. Minimnya jumlah soal pada rentang moderat (sukar dan mudah) menandakan bahwa instrumen ini kurang memiliki variasi soal di tingkat kemampuan rata-rata.

Secara keseluruhan, pola distribusi ini menunjukkan bahwa instrumen tes cenderung terpolarisasi. Keberadaan 50% soal dalam kategori sangat mudah menguntungkan untuk membangun kepercayaan diri siswa dan mengukur kompetensi dasar (*basic mastery*), namun kurang efektif dalam membedakan siswa berkemampuan sedang. Di sisi lain, adanya 30% soal sangat sukar berfungsi baik untuk menyeleksi siswa dengan kemampuan tinggi (*high achievers*).

Person & Item Separation dan Statistik Reliabilitas

Reliabilitas merupakan komponen krusial dalam penilaian kualitas suatu instrumen, terutama dalam konteks asesmen pembelajaran maupun penelitian pendidikan. Secara umum, reliabilitas didefinisikan sebagai derajat konsistensi atau kestabilan hasil yang diperoleh ketika suatu instrumen digunakan dalam pengukuran berulang terhadap kelompok responden yang memiliki karakteristik serupa (Mufrihah et al., 2025). Reliabilitas yang tinggi menunjukkan bahwa variasi skor lebih disebabkan oleh perbedaan kemampuan atau karakteristik peserta, bukan oleh kesalahan acak dalam pengukuran. Dengan kata lain, instrumen yang reliabel memungkinkan peneliti atau pendidik untuk membuat kesimpulan yang valid dan dapat dipercaya mengenai objek yang diukur (Arias, V. B., Ponce, F. P., & Almeida, 2016). Dalam pendekatan modern, reliabilitas tidak hanya dinilai berdasarkan konsistensi internal (seperti Cronbach's alpha), tetapi juga melalui estimasi dalam kerangka *Item Response Theory* (IRT) yang mempertimbangkan parameter individu dan butir secara simultan. Model Rasch, sebagai bagian dari IRT, memberikan indikator reliabilitas yang lebih mendalam seperti person separation reliability dan item separation reliability, yang merefleksikan kejelasan pemisahan antara kelompok responden berdasarkan kemampuan dan kestabilan item dalam mengukur populasi yang lebih luas (Zamani, B. E., & Ahmadi, 2016).

Tabel 3. Person & Item Separation dan Statistik Reliabilitas

Parameter	Measure (Mean)	SD	Separation	Reliability	Cronbach Alpha (KR-20)
Person (N=50)	1.75	1.3	0.95	0.47	0.61
Item (N=20)	-0.46	3.05	2.76	0.88	

Hasil analisis *Person & Item Separation* dan Statistik Reliabilitas dapat dilihat pada Tabel 3. Hasil analisis menunjukkan bahwa nilai *person separation* sebesar 0.95 mengindikasikan kemampuan instrumen untuk membedakan kemampuan individu (*person*) yang diuji masih tergolong rendah. Menurut Linacre (2020), nilai *separation* idealnya berada di atas 2.0, yang menunjukkan bahwa instrumen mampu mengelompokkan peserta ke dalam setidaknya tiga kelompok kemampuan yang berbeda. Nilai yang rendah seperti ini dapat terjadi karena rentang tingkat kesulitan item yang kurang luas atau populasi peserta yang homogen (Bond, T. G., & Fox, 2015). Oleh karena itu, diperlukan penambahan item dengan tingkat kesulitan yang lebih bervariasi, khususnya item yang lebih sukar mengingat rata-rata kemampuan responden (+1.75 logit) jauh lebih tinggi dari rata-rata kesulitan soal (-0.46 logit), untuk meningkatkan kemampuan instrumen dalam membedakan kemampuan siswa secara lebih tajam.

Sebaliknya, *item separation* mencapai nilai 2.76, yang menunjukkan bahwa tingkat kesulitan item dalam instrumen telah terdistribusi dengan cukup baik. *Separation* yang berada di atas 2.0 ini menandakan bahwa instrumen memiliki kualitas item yang stabil dan mampu mencakup berbagai tingkat kesulitan secara konsisten (Boone, W. J., Staver, J. R., & Yale, 2020). Nilai ini juga mengindikasikan bahwa *item reliability* sebesar 0.88 adalah baik, mengonfirmasi bahwa tingkat kesulitan item dapat diandalkan untuk mengukur kemampuan peserta dengan akurat. Dengan kata

lain, meskipun *person separation* memerlukan perbaikan, kualitas item yang tinggi menjadi keunggulan utama instrumen ini.

Menurut Pallant & Tennant (2007), *person separation* yang rendah dibandingkan *item separation* yang tinggi sering kali menunjukkan bahwa instrumen memerlukan penyesuaian untuk mencakup spektrum kemampuan peserta yang lebih luas (Pallant, J. F., & Tennant, 2007). Dalam konteks data ini, di mana *mean measure person* sangat tinggi, menambah item dengan tingkat kesulitan yang lebih tinggi (soal HOTS/Sukar) sangat diperlukan agar instrumen lebih sensitif terhadap peserta berkemampuan tinggi. Langkah ini juga dapat meningkatkan *person reliability*, yang dalam analisis ini bernilai 0.47 dan tergolong lemah.

Secara keseluruhan, hasil *separation* ini mencerminkan bahwa instrumen memiliki kualitas item yang baik, namun memerlukan penyesuaian untuk meningkatkan sensitivitas terhadap perbedaan kemampuan individu. Upaya untuk memperbaiki *separation* tidak hanya meningkatkan keakuratan pengukuran, tetapi juga validitas instrumen dalam berbagai konteks aplikas

Person reliability sebesar 0.47 mencerminkan kemampuan instrumen dalam membedakan peserta berdasarkan tingkat kemampuan mereka masih perlu ditingkatkan. Reliabilitas *person* dalam Rasch sering kali lebih rendah dibandingkan reliabilitas Cronbach karena Rasch lebih konservatif, namun lebih akurat dalam mencerminkan pengukuran yang bebas dari pengaruh data spesifik (Romdlon & Adi, 2025). Di sisi lain, *Item reliability* sebesar 0.88 menunjukkan konsistensi dan stabilitas yang tinggi dalam pengukuran tingkat kesulitan item. Ini mencerminkan bahwa instrumen ini mampu membedakan tingkat kesulitan item dengan baik di sepanjang rentang kemampuan peserta (Dharma & Meitiyani, 2022). Nilai *Cronbach Alpha* (KR-20) sebesar 0.61 mengindikasikan konsistensi internal yang berada pada kategori cukup, yang masih dapat diterima untuk tes tahap awal atau *pilot test*. Namun, untuk instrumen yang lebih matang, nilai ini perlu ditingkatkan hingga ≥0.8 dengan cara merevisi item yang tidak fit atau menambah jumlah butir soal.

KESIMPULAN DAN SARAN

Kesimpulan

Berdasarkan analisis data menggunakan Model Rasch terhadap instrumen penilaian Biologi pada materi Bioteknologi, diperoleh tiga kesimpulan utama terkait karakteristik psikometrik instrumen:

1. Instrumen memiliki validitas butir yang baik dengan *Item Reliability* yang tinggi (0.88) dan *Item Separation* yang produktif (2.76), menandakan bahwa butir soal konsisten dan mampu dikelompokkan berdasarkan tingkat kesulitannya. Namun, instrumen ini gagal dalam membedakan kemampuan siswa secara presisi, ditandai dengan *Person Reliability* yang lemah (0.47) dan *Person Separation* di bawah standar (0.95). Rendahnya reliabilitas responden ini disebabkan oleh ketidaksesuaian (*mismatch*) antara target ukur dengan alat ukur; kemampuan rata-rata siswa (+1.75 logit) jauh melampaui rata-rata kesulitan soal (-0.46 logit), sehingga instrumen terlalu mudah bagi responden (terjadi *ceiling effect*).
2. Distribusi tingkat kesukaran butir soal tidak proporsional dan terpolarisasi pada dua kutub ekstrem. Sebanyak 50% butir soal tergolong "Sangat Mudah" dan 30% tergolong "Sangat Sukar", sementara butir soal dengan kategori moderat ("Mudah" dan "Sukar") sangat minim (hanya 20%). Ketiadaan soal pada rentang moderat menyebabkan instrumen kehilangan sensitivitasnya untuk mengukur siswa dengan kemampuan rata-rata, yang berkontribusi langsung pada rendahnya nilai separasi individu.
3. Meskipun mayoritas item memenuhi kriteria validitas model Rasch, teridentifikasi tiga butir soal (S7, S14, dan S19) yang *misfit* (tumpang tindih atau *outlier*). Hal ini mengindikasikan adanya ambiguitas bahasa atau konsep pada butir tersebut yang membingungkan siswa. Selain itu, ditemukan 22% siswa dengan pola respons tidak wajar (*misfit person*), yang mengindikasikan adanya faktor tebakan (*guessing*) pada soal sukar atau kecerobohan (*carelessness*) pada soal mudah.

Saran

Saran Berdasarkan temuan di atas, beberapa rekomendasi untuk perbaikan instrumen dan evaluasi pembelajaran ke depan:

1. Guru perlu melakukan *re-targeting* instrumen dengan mengurangi porsi soal kategori "Sangat Mudah". Disarankan untuk menambah jumlah butir soal pada kategori moderat (logit -1.0 hingga +1.0) agar instrumen mampu membedakan kemampuan siswa secara lebih halus dan meningkatkan nilai *Person Reliability*.
2. Butir soal nomor S7, S14, dan S19 perlu ditinjau ulang dari segi konstruksi kalimat dan kesesuaian indikator. Jika revisi redaksional tidak memungkinkan, butir tersebut sebaiknya diganti (di-*drop*) untuk menjaga validitas konstruk pengukuran.
3. Mengingat rata-rata kemampuan siswa jauh di atas kesulitan soal (+1.75 logit), disarankan untuk meningkatkan level kognitif soal dari C1-C3 (LOTS/MOTS) menuju C4-C6 (HOTS). Hal ini diperlukan untuk menantang siswa berkemampuan tinggi dan menghindari penumpukan skor sempurna yang tidak informatif.
4. Rekomendasi bagi peneliti selanjutnya adalah disarankan menggunakan jumlah sampel yang lebih besar untuk menstabilkan estimasi *error* standar, serta membandingkan hasil analisis Rasch dengan analisis kualitatif (seperti *think-aloud protocol*) untuk memahami penyebab tingginya angka *person misfit*.

DAFTAR RUJUKAN

Agus Pambudi Dharma, & Meitiyani. (2022). Pelatihan Pengamatan Burung pada Pemuda Kampung Tabrik Desa Gekbrong Kabupaten Cianjur Dalam Mendukung Program Eduekowisata. *Jurnal Pengabdian Masyarakat Formosa*, *1*(2), 211–214. https://doi.org/10.55927/jpmf.v1i2.536

Ahmad, N. Q., Arthur, R., & Rahayu, W. (2025). *Pengembangan Instrumen Kemampuan Pemecahan Masalah Matematika Siswa Menggunakan Model Rasch*. *8*(1), 61–74.

Akram, W., Hussein, M. S. E., Ahmad, S., Mamat, M. N., & Ismail, N. E. (2015). Validation of the knowledge, attitude and perceived practice of asthma instrument among community pharmacists using Rasch analysis. *Saudi Pharmaceutical Journal*, *23*(5), 499–503. https://doi.org/10.1016/j.jsps.2015.01.011

Arias, V. B., Ponce, F. P., & Almeida, L. S. (2016). Validity and reliability of the school motivation scale using Rasch model. *European Journal of Psychological Assessment, 32*(3), 187–195. https://doi.org/. https://doi.org/10.1027/1015-5759/a000250

Bangi, U. K. M. (2022). Kesahan dan Kebolehpercayaan Instrumen Sikap Pelajar terhadap Pembelajaran Bahasa Arab (I-SPPBA) Menggunakan Model Pengukuran Rasch. *International Journal of West Asian Studies*, *14*, 98–110.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences (3rd ed.)*. https://doi.org/10.4324/9781315814578

Boone, W. J., Staver, J. R., & Yale, M. S. (2020). Rasch Analysis in the Human Sciences (2nd ed.). *Springer*. https://doi.org/https://doi.org/10.1007/978-94-017-9097-7

Dewi, K., Andriyani, K., & Model, R. (2018). Analisis Soal Pilihan Ganda dengan Mosdel Rasch. *Statistika*, *6*(1).

Farida dan Anna Musyarofah. (2021). *Validitas dan Reliabilitas dalam Analisis Butir Soal*. *I*(1), 34–44.

Fitraynsyah, A., & Hilmiyati, F. (2024). *Peran Tingkat Kesukaran dan Daya Pembeda dalam Analisis Butir Tes : Kajian Literatur untuk Pendidikan Menengah*. *1*(4), 252–262.

Ibrahim, N. L. (2025). *Analisis Literatur Penggunaan Model Rasch untuk Validasi Instrumen Asesmen Pendidikan : Tren , Temuan , dan Implikasi*. *3*(April).

Irwanto, J. (2025). *Jurnal Pendidikan Indonesia : Desain Instrumen Diagnostik Berbasis Model IRT Materi Sistem Persamaan Linier Perguruan Tinggi*. *5*(3). https://doi.org/10.59818/jpi.v5i3.1557

Noperi, H., Wiliyanti, V., & Yanti, F. A. (2024). *Natural : Jurnal Ilmiah Pendidikan IPA Validasi Instrumen Literasi Sains dengan Model Rasch untuk Socio-Scientific Issues*. *11*(2), 63–75. https://doi.org/10.30738/natural.v11i2.18891

Novriyanti, E., & Arthur, R. (2024). Analisis Kualitas Butir Soal Ujian Tengah Semester Biologi Umum Menggunakan Model Rasch. *JagoMIPA: Jurnal Pendidikan Matematika Dan IPA*, *4*(4), 718–733. https://doi.org/10.53299/jagomipa.v4i4.927

Pallant, J. F., & Tennant, A. (2007). *An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). British Journal of Clinical Psychology*. *46*(1), 1–18. https://doi.org/https://doi.org/10.1348/014466506X96931

Riskawati, Khaeruddin, Nirwana, Rifsa, T., Syasbila, A. F., & Syam, N. (2025). *Analisis Pemahaman Guru terhadap Model Pembelajaran Berbasis*. https://doi.org/10.30605/.XXX

Rivaldi, H. R. et al. (2025). AnalisisKualitas Butir Soal Asesmen Sumatif Geografi pada Materi Peta, Penginderaan Jauh,dan Sistem Informasi Geografis. *Jurnal Ilmu Pendidikan Dan Pembelajaran*, *04*(01), 102–113. https://doi.org/https://doi.org/10.58706/jipp

Romdlon, N., & Adi, M. (2025). *Analisis Kualitas Instrumen Literasi Media : Validitas dan Reliabilitas Menggunakan Model Rasch*. *10*(2), 323–336.

Ryan, T., Psikologi, F., Surabaya, U., Rungkut, K., Timur, J., Angella, S., Psikologi, F., Surabaya, U., Rungkut, K., Timur, J., Surya, R., Psikologi, F., Surabaya, U., Rungkut, K., & Timur, J. (2022). *Analisis Rasch Model Indonesia the International Personality Item Pool-Big Five Factor Marker s (IPIP-BFM-50)*. *10*(2), 297–317.

Simbolon, et. al. (2024). ANALISIS BUTIR SOAL MATEMATIKA PADA PENILAIAN AKHIR SEMESTER GANJIL MENGGUNAKAN RASCH MODEL KELAS X SMA NEGERI 2 BUNGURAN Program Studi Pendidikan Matematika , Universitas Maritim Raja Ali Haji , Tanjungpinang , Indonesia PENDAHULUAN Dunia pendidikan ditu. *Jumlahku*, *10*, 22–33.

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

Sumintono. (2016). Aplikasi Pemodelan Rasch pada asesmen pendidikan: Implementasi penilaian formatif (assessment for learning). *Makalah Dipresentasikan Dalam Kuliah Umum Pada Jurusan Statistika, Institut Teknologi Sepuluh November, Surabaya, 17 Maret 2016.*, *March*, 1–19. http://eprints.um.edu.my/15876/1/ITS_rasch_model_asesment_for_learning.pdf

Syafi'i et al. (2025). *EVALUASI PENDIDIKAN SEBAGAI DASAR PENGEMBANGAN INSTRUMEN PENILAIAN BERBASIS KOMPETENSI*. *1*(4), 1–12.

Tavakol, M., & Dennick, R. (2010). Psychometric evaluation of a knowledge based examination using Rasch analysis: An illustrative guide: AMEE Guide No. 72. *Medical Teacher*, *32*(12), 983–1000. https://doi.org/https://doi.org/10.3109/0142159X.2010.509414

Utami, A. P., Balqis, Z., Apriani, L., Rahmadani, I., & Dhea, R. (2025). Peran Evaluasi Pembelajaran Dalam Upaya Peningkatan Kualitas. *Al-Hasib: Manajemen Pendidikan Islam*, *02*(02), 434–446.

Zamani, B. E., & Ahmadi, M. (2016). Evaluation of e-learning readiness in higher education using the Rasch model. *Education and Information Technologies*, *21*(6), 1725–1742. https://doi.org/https://doi.org/10.1007/s10639-015-9418-6

Zhao, N. W., Mason, J. M., Blum, A. M., Kim, E. K., Young, V. V. N., Rosen, C. A., & Schneider, S. L. (2023). Using Item-Response Theory to Improve Interpretation of the Trans Woman Voice Questionnaire. *Laryngoscope*, *133*(5), 1197–1204. https://doi.org/10.1002/lary.30360