

ANALISIS KUALITAS BUTIR SOAL TES BIOLOGI MENGGUNAKAN RASCH MODEL

Intan Oktaviani¹, Pramesti Ayu Lestari², Erliana Putri Isnanto³, Khalishah Nabila Putri^{4*}, Muhammad Ilham Ardiansyah⁵, Suwarto⁶

¹Universitas Veteran Bangun Nusantara Sukoharjo, Sukoharjo, Indonesia

* Email: khalishahp978@gmail.com

Diterima: 27 Februari 2026

Direvisi: 04 Mei 2026

Publikasi: 15 Mei 2026

Abstract

This study aims to analyze the quality of a test instrument using the Rasch model with the assistance of the QUEST software. The analysis was conducted on ten test items developed as a learning evaluation tool. The results showed that only five items received responses from students and could be analyzed quantitatively. These five items demonstrated conformity with the Rasch model, as indicated by INFIT MNSQ values within the acceptable range and OUTFIT t values that met the item fit criteria. The analysis of item difficulty revealed that the instrument contained a range of difficulty levels, including easy, difficult, and very difficult items. However, the item reliability and person reliability values were classified as low, indicating that the consistency of the instrument and the stability of the measurement results were not yet optimal. In addition, several items did not function properly because they did not receive any student responses and therefore lacked empirical validity. Based on these findings, it can be concluded that the developed test instrument does not yet fully meet the criteria of a high-quality learning evaluation tool and still requires improvement, particularly in terms of item quality and instrument reliability.

Keywords: Rasch Model; Instrument Quality; Item Analysis; QUEST; Learning Evaluation

Abstrak

Penelitian ini bertujuan untuk menganalisis kualitas instrumen tes menggunakan model Rasch dengan bantuan perangkat lunak QUEST. Analisis dilakukan terhadap sepuluh butir soal yang dikembangkan sebagai alat evaluasi pembelajaran. Hasil penelitian menunjukkan bahwa hanya lima butir soal yang memperoleh respons peserta didik dan dapat dianalisis secara kuantitatif. Kelima butir tersebut menunjukkan kesesuaian dengan model Rasch berdasarkan nilai INFIT MNSQ yang berada dalam rentang penerimaan serta nilai OUTFIT t yang memenuhi kriteria kelulusan item. Analisis tingkat kesukaran menunjukkan bahwa instrumen memiliki variasi tingkat kesulitan, yaitu mudah, sulit, dan sangat sulit. Meskipun demikian, nilai reliabilitas item dan reliabilitas peserta tes tergolong rendah, yang mengindikasikan bahwa konsistensi instrumen dan stabilitas hasil pengukuran masih belum optimal. Selain itu, terdapat sejumlah butir soal yang tidak berfungsi karena tidak memperoleh respons peserta didik sehingga tidak memiliki validitas empiris. Berdasarkan temuan tersebut, dapat disimpulkan bahwa instrumen tes yang dikembangkan belum sepenuhnya memenuhi kriteria sebagai alat evaluasi pembelajaran yang berkualitas dan masih memerlukan perbaikan, terutama pada aspek kualitas butir soal dan reliabilitas instrumen.

Kata kunci: Model Rasch; Kualitas Instrumen; Analisis Butir Soal; QUEST; Evaluasi Pembelajaran

PENDAHULUAN

Evaluasi merupakan komponen penting dalam proses pembelajaran karena memberikan informasi mengenai pencapaian hasil belajar siswa secara berkesinambungan (Hilmiyati, 2024). Salah satu instrumen yang banyak digunakan dalam proses evaluasi adalah tes prestasi, yaitu alat ukur yang dirancang untuk

menilai sejauh mana siswa telah menguasai tujuan pembelajaran yang telah ditetapkan (Crocker & Algina, 1986; Suwarto, 2013). Kualitas tes prestasi sangat ditentukan oleh karakteristik butir soal, meliputi validitas, reliabilitas, tingkat kesukaran, daya pembeda, serta keefektifan distraktor sehingga diperlukan prosedur pengembangan tes yang sesuai dengan

kaidah pengukuran (Suwarto, 2021).

Dalam konteks pembelajaran Ilmu Pengetahuan Alam, khususnya mata pelajaran Biologi di SMA, tes hasil belajar berfungsi untuk mengukur kemampuan siswa dalam memahami konsep, menganalisis fenomena biologis, serta menerapkan pengetahuan dalam pemecahan masalah (Efendi, 2024). Fungsi ini sejalan dengan tuntutan kurikulum yang menekankan kemampuan berpikir kritis, literasi sains, dan pemahaman konseptual yang mendalam pada peserta didik. Oleh karena itu, penyusunan dan penggunaan tes Biologi yang bermutu menjadi aspek penting untuk memastikan bahwa hasil belajar yang diperoleh siswa benar-benar mencerminkan kemampuan yang ingin diukur. Instrumen yang tidak dianalisis kualitasnya berpotensi menghasilkan pengukuran yang bias dan tidak akurat (Zayrin et al., 2025).

Namun, pada praktiknya banyak guru masih menggunakan soal Biologi tanpa melalui proses analisis butir soal. Soal seringkali disusun berdasarkan pengalaman mengajar, diambil dari bank soal lama, atau diadaptasi dari sumber lain tanpa pemeriksaan terhadap kualitas item. Hal ini menyebabkan rendahnya validitas dan reliabilitas instrumen, distraktor yang tidak berfungsi, serta butir soal yang terlalu mudah atau terlalu sulit sehingga tidak mampu membedakan kemampuan siswa secara tepat (Darmawan et al., 2022). Kondisi ini menunjukkan adanya kebutuhan untuk menganalisis kualitas butir soal secara komprehensif agar instrumen evaluasi yang digunakan memiliki mutu yang baik.

Analisis kualitas butir soal dapat dilakukan melalui dua pendekatan utama, yaitu teori tes klasik (*Classical Test Theory/CTT*) dan teori tes modern (*Item Response Theory/IRT*) (Pratama, 2020). Teori tes klasik menawarkan prosedur analisis yang relatif sederhana melalui pengukuran validitas butir, reliabilitas tes, tingkat kesukaran, daya pembeda, serta

keefektifan distraktor. Meskipun demikian, hasil analisis dalam CTT cenderung bergantung pada karakteristik kelompok peserta tes.

Sebaliknya, teori tes modern, khususnya model Rasch, memberikan informasi yang lebih mendalam dan objektif mengenai kualitas butir soal karena estimasi parameter item dan kemampuan peserta tes bersifat lebih stabil serta tidak bergantung pada sampel tertentu. Model Rasch memungkinkan pemetaan kemampuan peserta didik dan tingkat kesukaran butir soal pada skala logit yang sama, sehingga dapat menggambarkan karakteristik instrumen secara lebih komprehensif. Oleh karena itu, penggunaan kedua pendekatan ini dapat menghasilkan evaluasi kualitas instrumen yang lebih menyeluruh dan akurat.

Penelitian ini bertujuan untuk menganalisis kualitas butir soal Biologi tingkat SMA berdasarkan teori tes klasik yang meliputi validitas butir, reliabilitas tes, tingkat kesukaran, daya pembeda, serta efektivitas distraktor. Selain itu, penelitian ini juga mengkaji kualitas instrumen menggunakan pendekatan teori tes modern melalui analisis model Rasch guna memperoleh gambaran karakteristik butir soal dan kemampuan peserta didik secara lebih mendalam. Hasil penelitian diharapkan dapat memberikan kontribusi bagi guru, pendidik, dan pengembang tes dalam meningkatkan mutu penyusunan serta penggunaan instrumen evaluasi pembelajaran Biologi di sekolah.

METODE PENELITIAN

Desain penelitian

Penelitian ini menggunakan desain penelitian deskriptif kuantitatif dengan pendekatan evaluatif. Penelitian bertujuan untuk menganalisis kualitas instrumen tes hasil belajar Biologi tingkat SMA menggunakan model Rasch. Pendekatan ini digunakan untuk mengetahui tingkat validitas, reliabilitas, tingkat kesukaran butir soal, serta kesesuaian item

terhadap model pengukuran sehingga diperoleh instrumen evaluasi yang layak digunakan dalam pembelajaran.

Objek/partisipan penelitian

Objek dalam penelitian ini adalah instrumen tes hasil belajar Biologi tingkat SMA yang terdiri atas beberapa butir soal pilihan ganda. Partisipan penelitian adalah peserta didik SMA di Kabupaten Sukoharjo yang menjadi responden dalam uji coba instrumen. Pemilihan partisipan dilakukan secara purposive sesuai dengan kebutuhan penelitian, yaitu peserta didik yang telah mempelajari materi Biologi yang diujikan.

Waktu dan lokasi penelitian

Penelitian yang dilakukan ini merupakan jenis penelitian deskriptif kuantitatif yang bertujuan untuk menganalisis butir soal hasil belajar biologi tingkat SMA. Kegiatan penelitian ini dilaksanakan selama 7 hari dari 1 Desember 2025 sampai pada tanggal 7 Desember 2025. Dan khusus pengumpulan data dilaksanakan tanggal 7 Desember 2025 di seluruh SMA di Sukoharjo.

Alat dan bahan

Peralatan yang digunakan berupa komputer/laptop, perangkat lunak (SPSS & Quest), google form, instrumen tes hasil belajar biologi, lembar validasi ahli, dan kisi-kisi instrumen tes.

Teknik pengumpulan data

Teknik pengumpulan data dalam penelitian ini dilakukan secara digital melalui beberapa tahapan yang sistematis. Instrumen tes hasil belajar Biologi disusun berdasarkan kompetensi dasar dan indikator pembelajaran yang tercantum dalam kurikulum. Tahap awal meliputi penyusunan kisi-kisi soal, penulisan butir soal, serta penelaahan awal untuk memastikan kesesuaian aspek materi, konstruksi, dan bahasa. Selanjutnya, draf instrumen divalidasi melalui *expert judgment*

oleh dosen atau guru Biologi yang kompeten untuk memperoleh masukan terkait kelayakan butir soal. Hasil validasi digunakan sebagai dasar perbaikan instrumen sebelum dilakukan uji coba. Instrumen yang telah direvisi kemudian disajikan dalam bentuk tes daring menggunakan Google Form dan diujicobakan kepada peserta didik SMA yang menjadi sampel penelitian. Respons peserta didik direkam secara otomatis dan diberi skor dikotomis, yaitu skor 1 untuk jawaban benar dan skor 0 untuk jawaban salah. Data yang terkumpul selanjutnya melalui proses *data cleaning* untuk memastikan tidak terdapat respons ganda, data tidak lengkap, maupun kesalahan entri, sehingga data siap untuk dianalisis. Prosedur pengumpulan data dilakukan melalui beberapa langkah, yaitu: (1) menyusun kisi-kisi dan butir soal berdasarkan kompetensi dasar; (2) melakukan validasi ahli terhadap instrumen; (3) merevisi instrumen sesuai saran validator; (4) menyebarkan instrumen melalui Google Form kepada peserta didik; (5) mengumpulkan hasil respons peserta didik; dan (6) melakukan data cleaning sebelum data dianalisis menggunakan model Rasch.

Teknik analisis data

Analisis data dalam penelitian ini menggunakan pendekatan teori tes modern melalui model Rasch dengan bantuan perangkat lunak QUEST. Model Rasch digunakan karena mampu memberikan estimasi parameter butir dan kemampuan peserta didik yang bersifat objektif serta tidak bergantung pada karakteristik kelompok sampel.

Analisis model Rasch meliputi beberapa tahapan, yaitu: (1) estimasi kesesuaian butir soal terhadap model Rasch melalui statistik *infit* dan *outfit mean square*; (2) estimasi tingkat kesukaran butir soal (*item difficulty*); (3) evaluasi kecocokan butir soal (*item fit*) berdasarkan kriteria model Rasch; dan (4) estimasi reliabilitas item dan person. Selain itu, dilakukan pemetaan kemampuan peserta didik dan tingkat kesukaran butir soal menggunakan Wright Map untuk

menggambarkan sebaran kemampuan dan karakteristik butir secara visual.

Butir soal dinyatakan memiliki kualitas yang baik apabila memenuhi kriteria kecocokan model Rasch, memiliki tingkat kesukaran yang proporsional, serta didukung oleh nilai reliabilitas item dan person yang memadai.

HASIL DAN PEMBAHASAN

Hasil analisis butir soal menggunakan model Rasch menunjukkan bahwa kualitas instrumen yang dikembangkan belum sepenuhnya memenuhi kriteria model yang diharapkan. Meskipun instrumen telah melalui tahapan penyusunan dan validasi oleh ahli, masih ditemukan sejumlah butir soal yang tidak sesuai dengan model Rasch, sebagaimana ditunjukkan oleh nilai *infit* dan *outfit mean square* di luar batas penerimaan. Temuan ini mengindikasikan bahwa proses validasi isi belum sepenuhnya menjamin kesesuaian empiris butir soal ketika diujicobakan kepada peserta didik.

Hasil perhitungan tersebut diinterpretasikan dengan menggunakan kriteria yang menyatakan bahwa jika indeks validasi butir soal $<0,4$ maka validitasnya rendah, antara $0,4-0,8$ maka validitasnya sedang, serta jika $>0,8$ mempunyai validitas tinggi (Hanna & Retnawati., 2022). Berlandaskan hasil analisis butir soal menggunakan model Rasch, instrumen penelitian menunjukkan bahwa dari sepuluh butir soal yang dianalisis, hanya lima butir soal, yaitu nomor 2, 4, 6, 8, dan 10, yang memperoleh respons peserta didik dan dapat dianalisis secara kuantitatif. Berdasarkan nilai daya pembeda (*discrimination*), butir soal nomor 2, 6, dan 8 berada pada kategori validitas sedang, sedangkan butir soal nomor 4 dan 10 berada pada kategori validitas rendah. Sementara itu, tidak ditemukan butir soal yang mencapai kategori validitas tinggi. Lima butir soal lainnya, yaitu nomor 1, 3, 5, 7, dan 9, tidak memperoleh

respons sama sekali (100% missing), sehingga tidak memiliki validitas empiris dan tidak dapat digunakan dalam proses pengukuran.

Oleh karena itu, secara keseluruhan instrumen tes belum sepenuhnya memenuhi kriteria sebagai alat evaluasi pembelajaran yang baik dan memerlukan revisi sebelum dapat dilanjutkan ke tahap penelitian berikutnya. Revisi meliputi perbaikan redaksi soal, pengoptimalan fungsi pengecoh, serta penggantian butir soal yang tidak berfungsi. Selanjutnya, uji coba instrumen dilaksanakan melalui media daring untuk memperoleh data respons peserta didik. Penelitian ini melibatkan 200 responden sebagai sampel uji coba instrumen. Analisis data dilakukan menggunakan perangkat lunak berbasis model Rasch untuk mengevaluasi pola tanggapan responden, kesesuaian item (item fit), tingkat kesukaran, daya pembeda, serta reliabilitas instrumen. Penggunaan model Rasch memungkinkan peneliti untuk menilai kualitas butir soal secara objektif dan mendeteksi butir yang tidak berfungsi dalam mengukur kemampuan peserta didik secara akurat.

1. Estimasi Kesesuaian Item

Menurut Setyawarno (2017), nilai INFIT MNSQ digunakan untuk membandingkan karakteristik setiap butir soal dengan kriteria model yang tercantum pada Tabel 1, guna mengetahui tingkat kesesuaian butir terhadap model Rasch dengan bantuan perangkat lunak QUEST.

Tabel 1. Kriteria Nilai INFIT MNSQ

Nilai INFIT MNSQ	Keterangan
$>1,33$	Tidak relevan dengan modal Rasch
$0,77-1,33$	Relevan dengan model Rasch
$<0,77$	Tidak relevan dengan modal Rasch

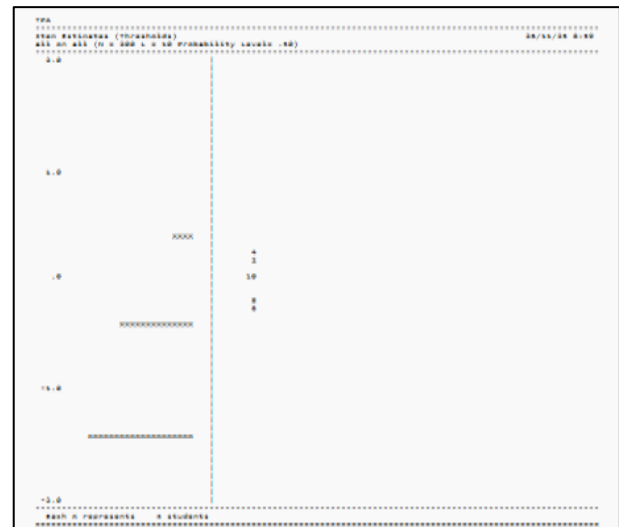
Berdasarkan kriteria kecocokan item terhadap model Rasch, item dinyatakan sesuai apabila memiliki nilai INFIT MNSQ berada pada rentang 0,77–1,33. Hasil analisis menggunakan program QUEST yang ditunjukkan pada Gambar 1 memperlihatkan bahwa tidak semua butir soal dapat dianalisis, karena beberapa item memiliki skor nol (*Item has zero score*).

Item Estimates (Thresholds) in Input Order									
all on all (N = 200 I = 10 Probability Level = .50)									
ITEM NAME	SCORE	RAWSCORE	INFIT	OUTFIT	INFIT	OUTFIT	INFIT	OUTFIT	INFIT
1	2	3	4	5	6	7	8	9	10
1 item 1	0	0	Item has zero score						
2 item 2	42	200	.13	.99	1.00	1.00	1.00	1.00	1.00
3 item 3	0	0	Item has zero score						
4 item 4	40	200	.15	1.02	1.03	1.03	1.03	1.03	1.03
5 item 5	0	0	Item has zero score						
6 item 6	60	200	-.10	.88	.87	1.00	1.00	1.00	1.00
7 item 7	0	0	Item has zero score						
8 item 8	57	200	-.10	1.03	1.02	1.02	1.02	1.02	1.02
9 item 9	0	0	Item has zero score						
10 item 10	40	200	.03	1.00	1.13	1.00	1.00	1.00	1.00
Mean			.00	1.00	1.01	1.01	1.01	1.01	1.01
SD			.13	.87	.89	1.11	1.11	1.11	1.11

Gambar 1. Rekapitulasi Item

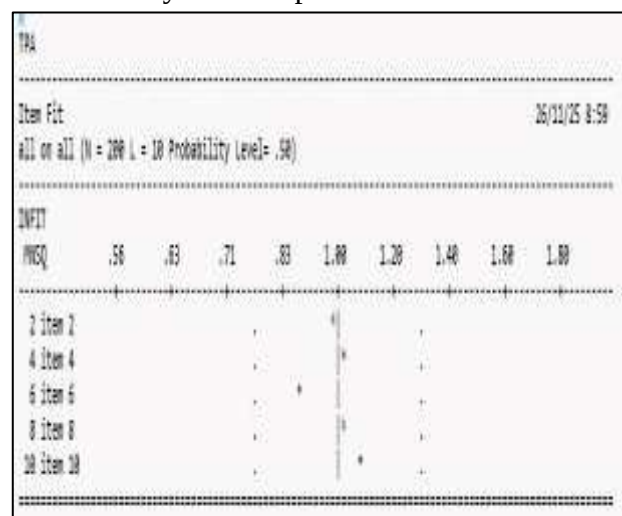
Adapun item yang memiliki nilai INFIT MNSQ dan dapat dievaluasi kesesuaiannya dengan model Rasch adalah item 2, item 4, item 6, item 8, dan item 10. Nilai INFIT MNSQ masing-masing item tersebut berturut-turut adalah 0,99; 1,02; 0,81; 1,01; dan 1,09. Seluruh nilai tersebut berada dalam rentang kriteria kecocokan 0,77–1,33, sehingga dapat disimpulkan bahwa item-item tersebut sesuai dengan model Rasch. Sementara itu, item lainnya tidak dapat dievaluasi kecocokannya karena tidak memiliki variasi skor responden. Selain menggunakan nilai INFIT MNSQ,

kecocokan butir soal terhadap model Rasch juga dapat ditinjau melalui peta kecocokan item (*item fit map*) sebagaimana disajikan pada Gambar 2.



Gambar 2. Fit map model Rasch

Berdasarkan Item Fit Map pada Gambar 2, garis vertikal menunjukkan nilai INFIT MNSQ = 1,00 sebagai nilai ideal kecocokan item terhadap model Rasch, dengan batas kecocokan berada pada rentang 0,77–1,33. Hasil analisis menunjukkan bahwa item 2, 4, 6, 8, dan 10 berada dalam rentang tersebut, sehingga dinyatakan sesuai dengan model Rasch. Sementara itu, item lainnya tidak dapat dianalisis karena memiliki skor nol (*item has zero score*), sehingga tidak dapat disimpulkan tingkat kecocokannya terhadap model Rasch.



Gambar 3. Sebaran tingkat kesukaraan Item

Gambar 3 menunjukkan hasil estimasi tingkat kesukaran butir soal berdasarkan nilai item estimate (threshold) menggunakan program QUEST. Berdasarkan hasil analisis, item 6 memiliki nilai item estimate paling tinggi sehingga merupakan item dengan tingkat kesukaran tertinggi, sedangkan item 10 memiliki nilai item estimate paling rendah sehingga tergolong item paling mudah bagi peserta tes. Klasifikasi tingkat kesukaran masing-masing item selanjutnya ditentukan berdasarkan kriteria yang tercantum pada Tabel 2 mengacu pada Setyawarno (2017).

Tabel 2. Kriteria Treshold

Nilai Treshold	Keterangan
$b > 2$	Sangat Sulit
$1 < b \leq 2$	Sulit
$-1 < b \leq 2$	Sedang
$-1 < b \leq -2$	Mudah
$b < -2$	Sangat Mudah

Tabel 3. Rekapitulasi tingkat kesukaran soal model Rasch

Item	Nilai Treshold	Interprestasi
1	-	-
2	0.19	sulit
3	-	-
4	0.25	sulit
5	-	-
6	-0.28	Mudah
7	-	-
8	-0.20	Mudah
9	-	-
10	1	Sangat sulit

Berdasarkan hasil analisis nilai threshold, dapat diketahui variasi tingkat kesulitan antarbutir soal. Item 6 dan item 8 memiliki nilai threshold negatif, masing-masing sebesar -0,28 dan -0,20, yang menunjukkan bahwa kedua item tersebut tergolong mudah, karena dapat dicapai oleh peserta didik dengan kemampuan relatif rendah. Item 2 dan item 4 memiliki nilai threshold positif namun masih berada pada rentang rendah, yaitu 0,19 dan 0,25, sehingga dikategorikan sulit, karena memerlukan

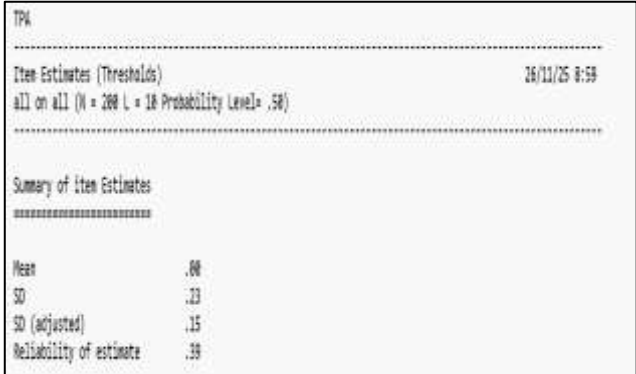
kemampuan peserta didik di atas rata-rata untuk mencapai kategori respon yang lebih tinggi.

Sementara itu, item 10 memiliki nilai threshold paling tinggi, yaitu sebesar 1, yang menunjukkan bahwa item tersebut tergolong sangat sulit dan hanya dapat dicapai oleh peserta didik dengan kemampuan tinggi. Adapun item 1, 3, 5, 7, dan 9 tidak memiliki nilai threshold, yang mengindikasikan bahwa butir tersebut bersifat dikotomi atau tidak menunjukkan peralihan kategori respon, sehingga tingkat kesulitannya tidak dapat ditentukan berdasarkan analisis threshold. Secara keseluruhan, tabel tersebut menunjukkan bahwa instrumen memiliki variasi tingkat kesulitan yang cukup beragam, mulai dari mudah hingga sangat sulit.

Tabel 4. Kriteria nilai estimate

Nilai Estimate	Keterangan
$> 1,00$	Kemampuan Tinggi
$-1,00 - 1,00$	Kemampuan Sedang
$< -1,00$	Kemampuan Rendah

Berdasarkan nilai Reliability of Case Estimate sebesar 0,25 yang berada dalam rentang -1,00 hingga 1,00, Gambar 4 menunjukkan bahwa kemampuan peserta tes berada pada kategori sedang.



Item Estimates (Thresholds)	
all on all (N = 200 L = 18 Probability Level = .50)	
Summary of Item Estimates	
Mean	.00
SD	.23
SD (adjusted)	.15
Reliability of estimate	.39

Gambar 4. Estimasi kemampuan responden

2. Estimasi item lolos (Fit)

Nilai t OUTFIT dalam program QUEST digunakan sebagai indikator untuk menentukan apakah suatu item dinyatakan lulus atau gagal. Item dikategorikan lulus apabila nilai t OUTFIT

$\leq 2,00$, sedangkan item dinyatakan gagal jika nilai t OUTFIT $\geq 2,00$ (Setyawarno, 2017). Gambar 1 menyajikan nilai t OUTFIT, sedangkan ringkasan komponen fit berdasarkan nilai tersebut disajikan pada Tabel 5.

Tabel 5. Rekapitulasi Item Fit

Item	Nilai OUTFIT t	Keterangan
1	-	-
2	0,0	Lulus
3	-	-
4	0,3	Lulus
5	-	-
6	-1,3	Lulus
7	-	-
8	0,2	Lulus
9	-	-
10	1,0	Lulus

Berdasarkan hasil analisis kesesuaian butir (item fit) menggunakan statistik OUTFIT t , diketahui bahwa sebagian besar item yang dapat dianalisis menunjukkan kesesuaian yang baik dengan model Rasch. Item dinyatakan lulus apabila memiliki nilai t OUTFIT $\leq 2,00$ (Setyawarno, 2017). Hasil pada Tabel 5 menunjukkan bahwa item 2, 4, 6, 8, dan 10 memiliki nilai t OUTFIT masing-masing sebesar 0,0; 0,3; -1,3; 0,2; dan 1,0. Seluruh nilai tersebut berada dalam rentang yang dapat diterima, sehingga kelima item tersebut dinyatakan lulus dan fit dengan model Rasch. Menurut Bond & Fox (2015), item yang fit menunjukkan bahwa pola respons peserta didik konsisten dengan kemampuan yang diukur, sehingga item mampu merepresentasikan konstruk yang sama secara valid dan objektif. Hal ini diperkuat oleh Linacre (2011) yang menyatakan bahwa statistik INFIT dan OUTFIT digunakan untuk mendeteksi kesesuaian respons peserta didik terhadap ekspektasi model Rasch.

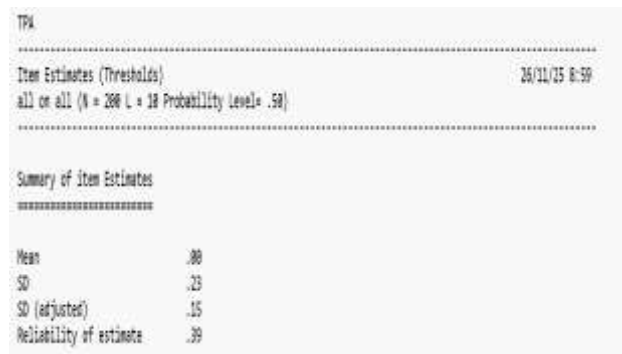
Apabila nilai OUTFIT berada pada rentang penerimaan, maka item dianggap mampu mengukur kemampuan peserta didik secara stabil dan tidak menunjukkan penyimpangan

respons yang signifikan. Temuan penelitian ini juga sejalan dengan penelitian Rizbudiani et al. (2021) yang menunjukkan bahwa item dengan nilai OUTFIT $t \leq 2,00$ dinyatakan sesuai dengan model Rasch karena pola jawaban responden masih berada dalam batas prediksi model. Selain itu, Rahim dan Haryanto (2021) menjelaskan bahwa item fit mengindikasikan adanya hubungan yang konsisten antara tingkat kemampuan peserta didik dan tingkat kesukaran item, sehingga kualitas pengukuran menjadi lebih akurat. Sementara itu, item 1, 3, 5, 7, dan 9 tidak menghasilkan nilai t OUTFIT sehingga tidak dapat dianalisis berdasarkan kriteria item fit. Kondisi ini diduga disebabkan oleh item yang tidak memperoleh variasi respons peserta didik (item has zero score), sehingga statistik OUTFIT t tidak dapat dihitung. Menurut Müller (2020), statistik fit pada model Rasch memerlukan adanya variasi pola respons agar estimasi kesesuaian item dapat dilakukan secara akurat.

Jika suatu item tidak memiliki variasi jawaban, maka item tersebut tidak mampu memberikan informasi pengukuran yang memadai. Selain itu, Tesio et al., (2024) menjelaskan bahwa keterbatasan variasi respons dapat menyebabkan item gagal mendeteksi perbedaan kemampuan peserta didik, sehingga item tidak berfungsi secara optimal dalam proses pengukuran. Dengan demikian, item yang tidak dapat dianalisis perlu direvisi, baik dari aspek redaksi, tingkat kesukaran, maupun fungsi pengecoh agar mampu menghasilkan variasi respons yang lebih baik pada uji coba berikutnya.. Oleh karena itu, kelayakan item-item tersebut tidak dapat ditentukan berdasarkan analisis kesesuaian model. Secara keseluruhan, hasil analisis item fit menunjukkan bahwa instrumen telah didominasi oleh item yang sesuai dengan model Rasch, sehingga secara umum instrumen layak digunakan untuk mengukur kemampuan peserta didik sesuai dengan tujuan penelitian.

3. Estimasi Realibilitas

Program QUEST dimanfaatkan untuk menghitung tingkat keandalan berdasarkan model Rasch. Gambar 4 dan Gambar 5 masing-masing menyajikan keandalan estimasi kasus serta keandalan estimasi item.



Summary of Item Estimates	
Mean	.00
SD	.23
SD (adjusted)	.15
Reliability of estimate	.39

Gambar 5. *Reliability of item Estimate*

Berdasarkan hasil analisis menggunakan program QUEST yang ditunjukkan pada Gambar 5, nilai *Reliability of Item Estimate* diperoleh sebesar 0,39. Dalam pemodelan Rasch, reliabilitas ini dikenal sebagai reliabilitas sampel. Mengacu pada kriteria yang dikemukakan oleh Susdelina et al. (2018), nilai reliabilitas < 0,67 termasuk dalam kategori lemah. Dengan demikian, nilai *Reliability of Item Estimate* sebesar 0,39 menunjukkan bahwa reliabilitas item masih tergolong lemah, yang mengindikasikan bahwa kemampuan item dalam membedakan karakteristik responden belum optimal dan kesesuaian item terhadap model Rasch masih terbatas. Sementara itu, nilai *Reliability of Case Estimate* atau reliabilitas peserta tes yang ditunjukkan pada Gambar 4 sebesar 0,25 juga berada pada kategori lemah. Nilai reliabilitas peserta tes yang rendah ini mencerminkan adanya ketidakkonsistenan dalam pola jawaban peserta.

Temuan tersebut sejalan dengan pendapat Ardiyanti (2017) yang menyatakan bahwa rendahnya reliabilitas responden dapat disebabkan oleh ketidaktelitian atau kecenderungan peserta tes dalam menjawab soal secara tidak konsisten. Apabila nilai *Reliability*

of Case Estimate berada pada kategori baik, maka pola jawaban peserta akan menunjukkan konsistensi yang lebih tinggi (Subali & Suyata, 2011; Dharma et al., 2024). Rendahnya nilai reliabilitas peserta tes mengindikasikan bahwa respon peserta masih belum stabil, yang berdampak pada rendahnya reliabilitas pengukuran. Meskipun demikian, hasil penelitian juga menunjukkan bahwa jumlah item soal tidak berkorelasi secara langsung dengan reliabilitas peserta tes. Temuan ini sejalan dengan penelitian Hakiki et al. (2018) yang menyatakan bahwa banyaknya item dalam suatu tes tidak selalu berpengaruh terhadap nilai reliabilitas pengukuran.

KESIMPULAN

Berdasarkan hasil analisis butir soal menggunakan model Rasch dengan bantuan perangkat lunak QUEST, dapat disimpulkan bahwa instrumen tes yang dikembangkan belum sepenuhnya memenuhi kriteria sebagai alat evaluasi pembelajaran yang berkualitas. Dari sepuluh butir soal yang dianalisis, hanya lima butir soal (item 2, 4, 6, 8, dan 10) yang memperoleh respons peserta didik dan dapat dianalisis secara kuantitatif. Kelima item tersebut menunjukkan kesesuaian yang baik dengan model Rasch berdasarkan nilai INFIT MNSQ yang berada dalam rentang penerimaan serta nilai OUTFIT t yang memenuhi kriteria kelulusan item.

Hasil analisis tingkat kesukaran menunjukkan bahwa instrumen memiliki variasi tingkat kesulitan yang beragam, mulai dari mudah hingga sangat sulit. Item 6 dan 8 tergolong mudah, item 2 dan 4 tergolong sulit, sedangkan item 10 tergolong sangat sulit. Variasi tingkat kesukaran ini menunjukkan bahwa instrumen memiliki potensi untuk mengukur kemampuan peserta didik pada berbagai tingkat, meskipun jumlah item yang dapat dianalisis masih terbatas. Sementara itu, lima item lainnya

tidak memperoleh respons sama sekali sehingga tidak memiliki validitas empiris dan tidak dapat digunakan dalam proses pengukuran. Ditinjau dari aspek reliabilitas, nilai *Reliability of Item Estimate* sebesar 0,39 dan *Reliability of Case Estimate* sebesar 0,25 menunjukkan bahwa reliabilitas instrumen maupun konsistensi jawaban peserta tes masih tergolong lemah. Rendahnya reliabilitas item mengindikasikan bahwa kemampuan butir soal dalam membedakan karakteristik responden belum optimal, sedangkan rendahnya reliabilitas peserta tes mencerminkan adanya ketidakkonsistenan dalam pola jawaban responden.

Kondisi ini menunjukkan bahwa hasil pengukuran belum sepenuhnya stabil dan akurat. Secara keseluruhan, meskipun sebagian item telah menunjukkan kesesuaian dengan model Rasch dan layak digunakan, instrumen tes secara umum masih memerlukan perbaikan sebelum digunakan pada tahap penelitian selanjutnya. Revisi instrumen perlu difokuskan pada perbaikan redaksi soal, pengoptimalan fungsi pengecoh, serta penggantian butir soal yang tidak berfungsi atau tidak memperoleh respons. Dengan demikian, diharapkan pada uji coba berikutnya instrumen dapat memiliki kualitas empiris yang lebih baik, baik dari sisi validitas maupun reliabilitas, sehingga mampu mengukur kemampuan peserta didik secara lebih akurat dan konsisten.

UCAPAN TERIMA KASIH

Penulis menyampaikan ucapan terima kasih kepada seluruh pihak yang telah memberikan bantuan, dukungan, dan kerja sama selama proses pelaksanaan penelitian hingga penyusunan naskah ini. Ucapan terima kasih secara khusus disampaikan kepada pihak sekolah yang telah memberikan izin dan fasilitas sehingga penelitian dapat terlaksana dengan baik, serta kepada para peserta didik yang telah

berpartisipasi secara aktif dan kooperatif dalam pengumpulan data penelitian. Penulis juga mengucapkan terima kasih kepada rekan sejawat dan dosen pembimbing yang telah memberikan masukan, saran, dan arahan yang konstruktif selama proses penelitian dan penulisan artikel ini. Semua kontribusi tersebut sangat berarti dalam meningkatkan kualitas penelitian dan penyempurnaan naskah. Penulis menyadari bahwa penelitian ini masih memiliki keterbatasan, sehingga kritik dan saran yang membangun sangat diharapkan untuk pengembangan penelitian selanjutnya.

DAFTAR PUSTAKA

- Ardiyanti, D. (2017). Aplikasi Model Rasch pada Pengembangan Skala Efikasi Diri dalam Pengambilan Keputusan Karir Siswa. *Jurnal Psikologi*, 43(3), 248. <https://doi.org/10.22146/jpsi.17801>
- Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.
- Darmawan, I. A. A., Zamsir, M. P., Widiastiti, A. A. I. P., Sos, S., Sugiharni, G. A. D., Amir, F. L., & Par, S. M. (2025). *Pengembangan Instrumen Penelitian*. Yogyakarta: Penerbit Naga Pustaka.
- Dharma, S., Jaelani, A. K., & Ariah, F. (2024). Desain Instrumen Pengukuran Literasi Lingkungan Siswa SMA dengan Menggunakan Pendekatan Model Rasch. *Jurnal Pendidikan Dan Kebudayaan*, 9(2), 218-236. <https://doi.org/10.24832/jpnk.v9i2.5126>
- Efendi, R. (2024). Analisis Hasil Uji Keterbacaan Petunjuk Praktikum Ekologi: Studi Kasus dengan Partisipasi 15 Mahasiswa. *Educatoria: Jurnal Ilmiah Ilmu Pendidikan*, 4(3), 101-108. <https://doi.org/10.36312/educatoria.v4i3.296>

- Hakiki, A. W., Fitri, A. R., & Agung, I. M. (2018). Analisis Properti Psikometri Subtes Merkaufgaben (ME) dengan Rasch Model. *Jurnal Psikologi*, 14(1), 40.
<https://doi.org/10.24014/jp.v14i1.4900>
- Hanna, W. F., & Retnawati, H. (2022). Analisis kualitas butir soal matematika menggunakan model rasch dengan bantuan software quest. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, 11(4), 3695-3704.
<http://dx.doi.org/10.24127/ajpm.v11i4.5908>
- Hilmiyati, F. (2024). Integration of Cognitive Technology in Learning Assessment and Evaluation. *Al-Hijr: Journal of Adulearn World*, 3(2).
<http://dx.doi.org/10.55849/alhijr.v3i2.668>
- Linacre, J. M. (2011). A User's Guide to WINSTEPS MINISTEP Rasch-Model Computer Programs. *winsteps.com. Chapter, 18*, 567-573.
- Müller, M. (2020). Item fit statistics for Rasch analysis: can we trust them?. *Journal of Statistical Distributions and Applications*, 7(1), 5.
<https://doi.org/10.1186/s40488-020-00108-7>
- Perdana, S. A. (2018). Analisis kualitas instrumen pengukuran pemahaman konsep persamaan kuadrat melalui teori tes klasik dan rasch model. *Jurnal Kiprah*, 6(1), 41-48.
<https://doi.org/10.31629/kiprah.v6i1.574>
- Pratama, D. (2020). Analisis kualitas tes buatan guru melalui pendekatan item response theory (IRT) model rasch. *Tarbawy: Jurnal Pendidikan Islam*, 7(1), 61-70.
<https://doi.org/10.32923/tarbawy.v7i1.1187>
- Rizbudiani, A. D., Jaedun, A., Rahim, A., & Nurrahman, A. (2021). Rasch model item response theory (IRT) to analyze the quality of mathematics final semester exam test on system of linear equations in two variables (SLETV). *Al-Jabar: Jurnal Pendidikan Matematika*, 12(2), 399-412.
<https://doi.org/10.24042/ajpm.v12i2.9939>
- Rahim, A., & Haryanto, H. (2021). Implementation of Item Response Theory (IRT) Rasch model in quality analysis of final exam tests in Mathematics. *Journal of Research and Educational Research Evaluation*, 10(2), 57-65.
<https://doi.org/10.15294/jere.v10i2.51802>
- Setyawarno, D. (2017). *Use of Aplication of Software Itean (Item and Test Analysis) toAnalysis of Multiple-Choice Item Based upon Classical Test Theory*. Universitas Negeri Yogyakarta.
- Subali, B., & Suyata, P. (2011). *Panduan analisis data pengukuran pendidikan untuk memperoleh bukti empirik kesahihan menggunakan program Quest*. Universitas Negeri Yogyakarta: Lembaga Penelitian dan Pengabdian pada Masyarakat.
- Suwarto, S. (2013). *Pengembangan tes prestasi belajar*. Yogyakarta: Pustaka Pelajar.
- Suwarto, S. (2021). *Analisis butir soal*. Yogyakarta: Pustaka Pelajar.
- Tesio, L., Caronni, A., Simone, A., Kumbhare, D., & Scarano, S. (2024). Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment. *Disability and Rehabilitation*, 46(3), 604-617.
<https://doi.org/10.1080/09638288.2023.2169772>
- Zayrin, A. A., Nupus, H., Maizia, K. K., Marsela, S., Hidayatullah, R., & Harmonedi, H. (2025). Analisis Instrumen Penelitian Pendidikan (Uji Validitas Dan Relibilitas Instrumen Penelitian). *QOSIM: Jurnal Pendidikan Sosial & Humaniora*, 3(2), 780-789.
<https://doi.org/10.61104/jq.v3i2.1070>