



Analisis kualitas soal penilaian tengah semester biologi dengan teori tes klasik dan *Rasch* model

Damasus Nosi Resi^{a, 1, *}

^a Program Studi Pendidikan Biologi, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Katolik Widya Mandira, Kupang, Nusa Tenggara Timur, 85225, Indonesia.

¹ nosiresi@gmail.com *

* Corresponding author.

INFORMASI ARTIKEL

Lini Masa Artikel

Draft diterima : 2024-06-12
Revisi diterima : 2024-08-22
Diterbitkan : 2024-10-23

Keywords

Classical test theory;
Item analysis;
Rasch Model;

ABSTRAK

Tujuan penelitian ini untuk mengetahui kualitas soal Penilaian Tengah Semester ganjil kelas X mata pelajaran Biologi tahun pelajaran 2023/2024. Penelitian ini menggunakan metode deskriptif kuantitatif yang melibatkan 30 siswa dengan 30 soal yang diujikan. Data dianalisis dengan teori tes klasik yang diolah menggunakan Microsoft Excel dan *Rasch* model yang diolah dengan program Winstep. Kualitas item soal yang diuji meliputi tingkat kesukaran, daya pembeda, efektivitas distraktor, validitas dan reliabilitas. Hasil analisis CTT menunjukkan tingkat kesukaran terdapat dua kategori yaitu soal kategori mudah (13,3%) dan sedang (86,7%). Daya pembeda terdapat empat kategori yaitu rendah, cukup, tinggi dan sangat tinggi. Efektifitas distraktor berfungsi sebanyak 16,7% sedangkan yang tidak berfungsi 83,3%. Validitas butir soal yang valid terdapat 20% dan tidak valid 80%. Sedangkan tingkat reliabilitas dikategorikan tidak reliabel dengan nilai $r_{11} = 0,079$ (sangat rendah). Hasil analisis dengan IRT menunjukkan tingkat kesukaran soal terdapat tiga kategori yaitu mudah (20%), sedang (66,7%) dan sulit (13,3%). Daya pembeda hanya dikategorikan rendah. Efektifitas distraktor yang berfungsi sebanyak 43,3% dan tidak berfungsi 56,7%. Validitas butir soal tergolong valid 63,7% dan tidak valid 36,3. Reliabilitas person sebesar 0,03 dengan kategori lemah sedangkan reliabilitas item sebesar 0,30 dengan kategori lemah. Penelitian ini memberikan wawasan tentang pentingnya menggunakan metodologi yang tepat dalam analisis soal, baik menggunakan teori tes klasik maupun model *Rasch*, yang dapat membantu dalam pembuatan instrumen evaluasi di masa depan.

ABSTRACT

This research aims to determine the quality of odd semester assessment questions for class X in Biology for the 2023/2024 academic year. This research used a quantitative descriptive method involving 30 students with 30 questions being tested. Data were analyzed using classical test theory and processed using Microsoft Excel and *Rasch* models, which were processed using the Winstep program. The quality of the item questions tested includes difficulty level, distinguishing power, distractor effectiveness, validity, and reliability. The results of the CTT analysis show two categories of difficulty, namely easy (13.3%) and medium (86.7%) questions. There are four categories of differentiating power: low, sufficient, high, and very high. The effectiveness of functioning distractors was 16.7%, while non-functional ones were 83.3%. The validity of valid questions is 20%, and invalid is 80%. Meanwhile, the level of reliability is categorized as unreliable, with a value of $r_{11} = 0.079$ (very low). The results of the IRT analysis show that the difficulty level of the questions is in three categories, namely easy (20%), medium (66.7%), and difficult (13.3%). Differentiating power is only categorized as low. The effectiveness of functioning distractors was 43.3%, and non-functioning was 56.7%. The validity of the questions is classified as valid 63.7% and invalid 36.3. Person reliability is 0.03 in the weak category, while item reliability is 0.30 in the weak category. This study provides insight into the importance of using appropriate methodology for problem analysis, using both classical test theory and the *Rasch* model, which can help create evaluation instruments in the future.

Cara Sitasi Artikel Ini (APA Style):

Rosi, D. N. (2024). Analisis kualitas soal penilaian tengah semester biologi dengan teori tes klasik dan *Rasch* model. *Bio-Pedagogi*. 13(2), 63-77. <https://dx.doi.org/10.20961/bio-pedagogi.v13i2.88075>.

Artikel ini dapat diakses secara bebas dengan lisensi [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/).



PENDAHULUAN

Pendidikan berkualitas tidak terlepas dari proses pelaksanaan kegiatan belajar mengajar yang disertai dengan evaluasi yang baik. Evaluasi merupakan proses untuk menentukan sejauh mana, dalam hal apa, dan bagaimana tujuan pendidikan sudah tercapai ([Lestari et al., 2022](#)). Evaluasi pembelajaran bertujuan untuk menilai sejauh mana kemampuan siswa dalam memahami materi yang telah disampaikan oleh guru ([Maulida et al., 2023](#)). Evaluasi hasil belajar pada umumnya mencakup bagaimana siswa memperoleh hasil kognitif, psikomotorik, dan afektif dalam jangka pendek, jangka menengah dan jangka panjang yang mana mencakup semua aspek pendidikan seperti tujuan, sasaran, dan isi program. Tujuan evaluasi proses adalah untuk menilai seberapa efektif dan efisien kegiatan pembelajaran sehingga dapat disempurnakan ([Alfarisa & Purnama, 2019](#)).

Keberhasilan pembelajaran dapat dilihat dari hasil evaluasi pembelajaran. Tes merupakan teknik evaluasi pembelajaran dimana siswa diminta untuk menyelesaikan berbagai butir atau tugas yang harus dikerjakan untuk menghasilkan nilai tentang perilaku siswa ([Purwaningsih, 2022](#)). Tes merupakan alat untuk mengukur ketercapaian tujuan pembelajaran yang valid dan reliabel serta mengetahui hasil belajar siswa setelah mengikuti proses pembelajaran dalam jangka waktu dan satuan pendidikan tertentu ([Prasetya & Pratama, 2023](#)).

Pada jenjang sekolah menengah atas salah satu cara untuk menilai hasil belajar siswa adalah dengan Penilaian Tengah Semester (PTS). PTS digunakan oleh guru untuk mengetahui seberapa baik siswa mencapai kompetensi mereka pada pertengahan satu semester pembelajaran. Ada berbagai jenis soal untuk mengevaluasi kompetensi siswa, seperti soal pilihan ganda, uraian panjang, atau jawaban singkat. Guru biasanya memilih pilihan ganda karena lebih mudah untuk menghitung skor, mengoreksi atau memeriksa jawaban siswa. Apalagi dengan adanya program atau aplikasi komputer saat ini, hal ini lebih mudah dilakukan. Selain itu, materi yang dapat dievaluasi dengan pilihan ganda tidak hanya lebih banyak daripada soal uraian, tetapi juga lebih luas ([Eliza et al., 2022](#)).

Analisis soal adalah proses yang sistematis untuk mengevaluasi kualitas setiap komponen soal. Hasil dari analisis setiap item soal dapat menghasilkan informasi khusus serta untuk menemukan soal yang salah atau tidak berfungsi. Analisis soal penting digunakan untuk menguji kriteria instrumen penilaian hasil belajar dan untuk mengevaluasi kemampuan siswa ([Sari et al., 2022](#)). Tujuan analisis soal adalah untuk mengetahui kualitas. Analisis soal mencakup kegiatan menelaah dan mengkaji setiap item soal untuk menghasilkan soal yang berkualitas tinggi sehingga soal yang tidak baik juga dapat diperbaiki, direvisi atau dibuang ([Handayani, 2022](#)).

Soal dikatakan bermutu jika ia dapat memberikan informasi sesuai dengan tujuannya, seperti mengetahui apakah siswa memiliki kemampuan tinggi atau rendah terhadap materi yang diujikan serta dapat mengetahui kualitas soal terkait tingkat kesukarannya dan validitasnya. Analisis butir soal sangat penting untuk meningkatkan proses belajar mengajar karena guru dapat menganalisis materi pelajaran apa yang sulit dan apa yang mudah dipahami siswa melalui soal yang telah diberikan kepada mereka ([Handayani, 2022](#)).

Analisis soal penilaian tengah semester sangat penting untuk memperbaiki dan meningkatkan kualitas soal yang akan diujikan pada periode selanjutnya. Soal-soal ini dianalisis untuk mengidentifikasi soal yang baik dan buruk, dan soal-soal yang tidak baik dapat digunakan sebagai ukuran dan acuan untuk menyelesaikan ujian pada periode selanjutnya. Soal yang buruk yang masih dapat diperbaiki sehingga dapat disimpan di bank dan dapat digunakan kembali. Soal yang sangat buruk sebaiknya dibuang. Salah satu cara identifikasi kualitas soal adalah dengan memeriksa validitas, reliabilitas, kepraktisan, analisis butir soal (tingkat kesukaran dan daya beda), analisis pengecoh, analisis homogenitas soal dan efektivitas fungsi opsi.

Analisis butir soal dapat dilakukan dengan cara klasik yang dikenal dengan teori tes klasik atau *Classical Test Theory* (CTT) maupun dengan cara modern yang dikenal dengan teori respons butir atau *Item Response Theory* (IRT). Teori tes klasik didasarkan pada skor amatan yang merupakan penjumlahan dari skor sebenarnya dan skor kesalahan pengukuran dalam suatu analisis instrumen soal ([Skyttberg et al., 2023](#)).

Teori tes klasik masih dirasa memiliki beberapa kelemahan, misalnya tingkat kesukaran serta daya beda butir soal tergantung pada kelompok peserta yang mengerjakan sehingga seorang ahli matematika dari Denmark bernama Dr. Georg Rasch berhasil menemukan sebuah teori baru yang modern untuk mengatasi kelemahan-kelemahan cara klasik yang dinamakan *Item Response Theory* (IRT). IRT dianggap dapat merepresentasikan kemampuan setiap individu pada berbagai tingkat kemampuan siswa yang mengikuti ujian tersebut ([Fauziana & Wulansari, 2021](#)).

IRT dapat digunakan sebagai alat analisis untuk mengukur kemampuan siswa dalam konteks penilaian sains praktis dalam ujian berbasis kinerja tinggi. Model *Rasch* adalah model pengukuran yang digunakan untuk mengonversi data ordinal menjadi skala interval, sehingga memungkinkan perbandingan yang lebih akurat antara kemampuan siswa dan tingkat kesulitan soal ujian. Dengan menerapkan IRT, peneliti dapat mengukur sejauh mana siswa mampu menjawab pertanyaan-pertanyaan praktis sains yang diajukan dalam ujian. Model ini juga dapat membantu dalam mengidentifikasi aspek-aspek spesifik dari keterampilan praktis sains yang mungkin perlu ditingkatkan dalam kurikulum atau metode pengajarannya ([Myrzagaliyeva et al., 2024](#)).

Guru dan calon guru dituntut mahir dalam merencanakan proses pembelajaran, menentukan tujuan pembelajaran berdasarkan indikator, menentukan bahan pelajaran secara menyeluruh, memilih, dan menetapkan metode pembelajaran yang sesuai dengan materi, dan menyiapkan alat dan media pembelajaran ([Handayani, 2022](#)). Namun banyak guru atau calon guru yang belum tahu bagaimana menganalisis penilaian hasil belajar siswa dikarenakan minimnya pengetahuan dan pengalaman dalam menganalisis data nilai secara mendalam ([Andayani et al., 2017](#)).

Sebagian guru hanya tahu cara memasukkan nilai ke dalam daftar nilai, tetapi belum memahami bagaimana seharusnya mengolah dan menginterpretasikan data tersebut untuk kepentingan evaluasi dan perencanaan pembelajaran selanjutnya, padahal penilaian hasil belajar sangat penting sebagai bahan evaluasi yang sangat tepat untuk mengukur ketercapaian kompetensi siswa dan memberikan refleksi bagi guru untuk melanjutkan dan menyesuaikan materi selanjutnya ([Andayani et al., 2017](#)).

Oleh karena itu, semua guru dan calon guru harus memiliki kemampuan dan keterampilan untuk melakukan penilaian dan evaluasi. Perlu latihan dan pengalaman lapangan yang cukup agar dapat menguasai kemampuan ini, sehingga ketika guru atau calon guru telah memiliki kemampuan untuk mengevaluasi hasil belajar siswa, mereka bukan hanya membantu dalam proses belajar mengajar tetapi juga membantu dalam penelitian ilmiah, yaitu membuat alat ukur yang valid dan dapat diandalkan melalui analisis yang dilakukan.

Analisis ini bertujuan mendeteksi kualitas butir soal dari instrumen soal PTS Biologi kelas X SMA yang mencakup validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas distraktor sebagai informasi diagnostik untuk menentukan apakah siswa memahami materi yang telah dipelajari dan untuk meningkatkan pengelolaan pelaksanaan pembelajaran dan kualitas pertanyaan yang dibuat oleh guru dalam instrumen penilaian dan evaluasi.

METODE

Penelitian ini dilakukan dengan pendekatan deskriptif kuantitatif dimana metode analisis soal yang digunakan dalam kajian ini adalah teori tes klasik atau *Classical Test Theory* (CTT) menggunakan aplikasi *Microsoft Excel* dan *Item Response Theory* (IRT) dengan model *Rasch* melalui aplikasi *Winstep*. Metode CTT dipilih karena dianggap lebih mudah sedangkan pemilihan metode IRT karena metode ini memungkinkan analisis data secara lebih mendalam, bahkan pada tingkat per siswa maupun per item soal. IRT mampu menutupi kelemahan yang ditemukan dalam analisis menggunakan CTT karena menghasilkan kesimpulan lebih akurat dan bermakna atas data yang terkumpul ([Subali et al., 2021](#)). Instrumen tes yang dilakukan analisis merupakan soal penilaian tengah semester (PTS) semester ganjil tahun pelajaran 2023/2024 Kelas X pada mata pelajaran Biologi. Instrumen ini terdiri dari 30 butir soal pilihan ganda yang diujikan kepada 30 siswa.

HASIL DAN PEMBAHASAN

Hasil analisis soal PTS, yang meliputi tingkat kesukaran (*difficulty indeks*), daya pembeda, efektivitas distraktor, validitas dan reliabilitas menggunakan metode CTT dijelaskan dalam Lampiran 1. Semakin kecil nilai DI maka soal semakin sulit. Sedangkan nilai DP semakin kecil maka semakin rendah daya pembeda soal tersebut. Sebagian besar distraktor soal sudah efektif tetapi ada yang belum memiliki distraktor efektif.

Hasil menunjukkan terdapat variasi data yang sangat mencolok dimana terdapat soal dengan nilai DI sedang dan DP rendah sehingga tergolong soal tidak valid yaitu pada soal nomor 3, 4, 6, 8, 12, 14, 15, 16, 17, 18, 19, 21, 23, 24, dan 26. Kemudian soal dengan DI sedang dan DP cukup sehingga tergolong soal tidak valid terdapat pada soal nomor 1, 2, 10, 11, 22 dan 27. Ada juga soal dengan ID mudah dan DP rendah sehingga tergolong soal tidak valid yaitu pada soal nomor 15 dan 29. Tetapi ada juga soal dengan ID mudah dan DP cukup sehingga tergolong soal tidak valid terdapat pada soal nomor 5 dan 9. Sedangkan soal dengan ID sedang dan DP tinggi sehingga tergolong valid yaitu pada soal nomor 13, 20, 25, 28, dan 30. Terdapat pula soal dengan DI mudah dan DP Sangat tinggi sehingga tergolong valid yaitu pada soal nomor 7. Berikut ini disajikan ringkasan kualitas butir soal PTS Biologi kelas X dengan menggunakan CTT.

Tabel 1. Ringkasan Persentase Kualitas Butir Soal

No.	Indikator							
	DI	%	DP	%	Distraktor	%	Validitas	%
1	Mudah	13,3	Sangat Rendah	0	Berfungsi ($\geq 5\%$)	16,7	Valid	20
2	Sedang	86,7	Rendah	53,3	Tidak Berfungsi ($< 5\%$)	83,3	Tidak Valid	80
3	Sulit	0	Cukup	26,7				
4			Tinggi	16,7				
5			Sangat Tinggi	3,3				

Tabel 1 menjelaskan bahwa persentase tingkat kesukaran (DI) paling tinggi adalah pada soal dengan kategori sedang yaitu sebesar 86,7%. Persentase tertinggi pada indikator daya pembeda (DP) adalah pada soal dengan kategori daya pembedanya rendah yaitu sebesar 53,3%. Indikator keefektifan distraktor yaitu 83,3% opsi pengecoh adalah tidak berfungsi, artinya distraktor/ pengecoh tersebut tidak dipilih $\geq 5\%$ oleh siswa-siswa peserta ujian. Validitas soal menunjukkan 80% butir soal adalah tidak valid. Secara keseluruhan instrumen soal PTS ini adalah tidak reliabel yaitu dengan nilai *Cronbach Alpha* (r_{11}) = 0,079 dan termasuk kategori sangat rendah.

Tingkat kesukaran suatu butir soal adalah proporsi antara siswa yang dapat menjawab benar soal tersebut dengan jumlah seluruh siswa yang mengikuti tes/ujian. Soal yang berkualitas baik adalah soal yang tidak terlalu mudah namun juga tidak terlalu sulit. Dengan demikian soal yang baik adalah soal yang memiliki tingkat kesukaran dengan kategori sedang sampai tinggi sehingga soal dibuat memiliki daya kesukaran (Stano et al., 2024). Kategori tingkat kesukaran disajikan pada **Tabel 2**.

Tabel 2. Kategori Tingkat Kesukaran Soal

Koefisien	Kategori Tingkat Kesukaran
$p < 0,3$	Sukar
$0,3 \leq p \leq 0,7$	Sedang
$p > 0,7$	Mudah

Hasil analisis menunjukkan bahwa tidak ada soal yang berkategori sulit. Soal yang berkategori sedang mendominasi sebagian besar soal. Berarti instrumen ini belum memenuhi proporsi perbandingan yang baik, proporsi perbandingan 3-5-2 atau 30% butir soal dengan kategori mudah, 50% butir soal dengan kategori sedang, dan 20% butir soal dengan kategori sukar.

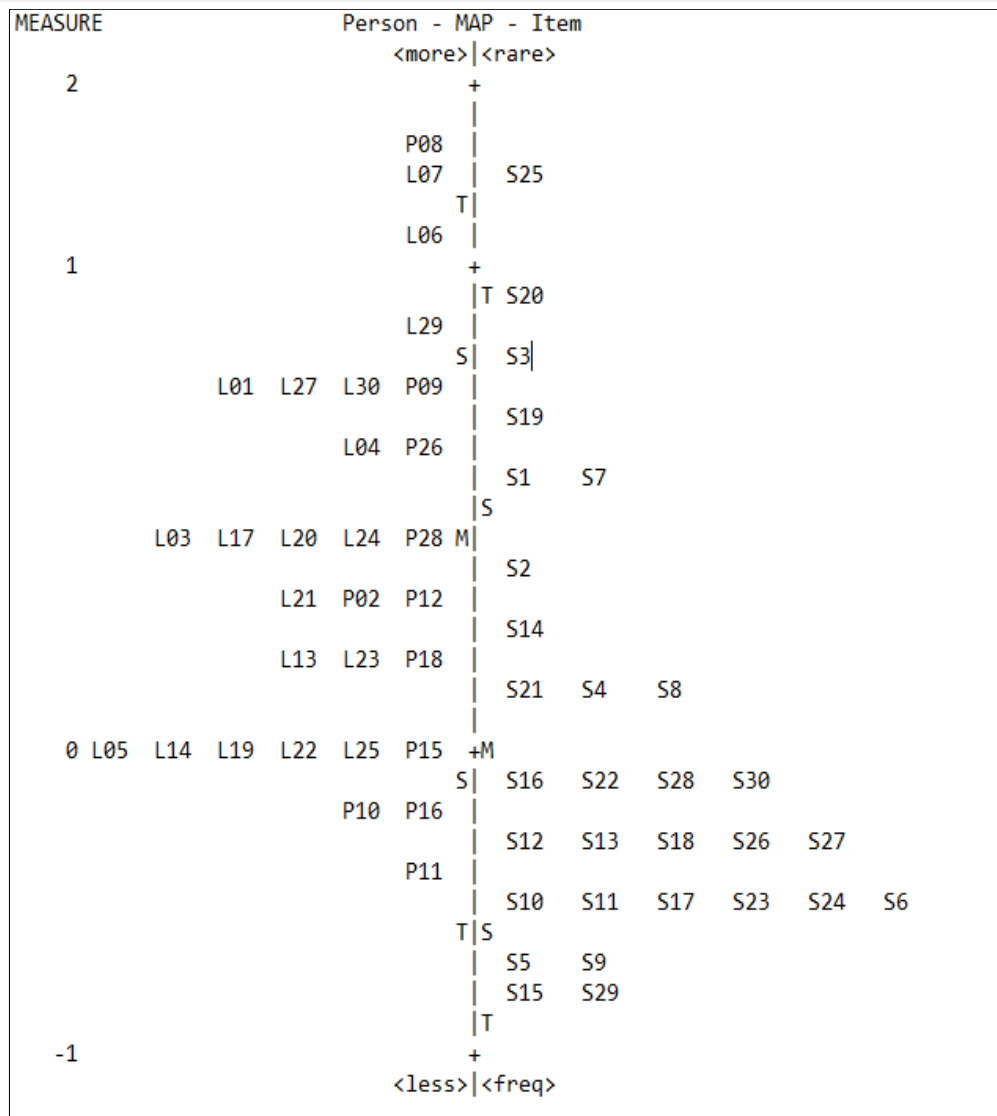
Analisis daya pembeda dilakukan untuk mengetahui kemampuan soal dalam membedakan siswa yang memiliki kemampuan tinggi dengan siswa yang memiliki kemampuan rendah. Daya pembeda berfungsi untuk mendeteksi perbedaan individual di antara para peserta tes sekecil-kecilnya ([Erfan et al., 2020](#)). Hasil analisis menunjukkan bahwa soal memiliki daya pembeda sangat rendah, rendah, cukup, tinggi dan sangat tinggi dengan proporsi tidak merata. Namun soal dengan kategori rendah menunjukkan proporsi yang agak seimbang. Hal ini berarti tidak sesuai bahwa soal dapat dikatakan telah memenuhi kriteria apabila terletak pada kategori sedang, tinggi, dan sangat tinggi. Sedangkan soal dikatakan tidak memenuhi kriteria apabila masuk kategori rendah dan sangat rendah. Soal yang tidak memenuhi kriteria berarti soal tersebut belum mampu membedakan siswa yang memiliki kemampuan tinggi dengan siswa yang memiliki kemampuan rendah.

Suatu distraktor (pengecoh) dikatakan efektif jika dipilih oleh minimal dipilih oleh 5% peserta tes. Keefektifan distraktor dikatakan berfungsi jika opsi pengecoh tersebut banyak dipilih oleh peserta tes dari kelompok bawah. Namun jika banyak kelompok atas yang memilih opsi pengecoh tersebut maka dikatakan keefektifan pengecoh tidak berfungsi ([Misri et al., 2021](#)). Analisa juga menunjukkan bahwa hampir sebagian besar pengecoh tidak berfungsi atau keefektifannya kurang signifikan yaitu soal nomor 1, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 20, 21, 22, 23, 25, 26, 27, 28, 29 dan 30. Pada butir-butir soal ini terdapat opsi jawaban yang dijawab siswa kurang dari 5%. Hal ini menunjukkan bahwa pengecoh terlalu mencolok perbedaannya bila dibandingkan pilihan jawaban yang lainnya sehingga kurang menimbulkan ketertarikan siswa untuk memilih pengecoh tersebut.

Validitas menunjukkan seberapa tepat dan cermat suatu butir soal dapat mengukur apa yang akan diukur, kaitannya dengan pembelajaran tentunya hal ini adalah hasil belajar siswa. Soal dikatakan valid jika mampu mengukur objek tersebut sesuai kriteria yang telah ditentukan. Jika validitas tes rendah akan menghasilkan data yang tidak relevan dengan tujuan pengukuran. Validitas butir soal dilihat berdasarkan perhitungan korelasi *product moment*. Soal dikatakan valid jika koefisien korelasi $r_{xy} > r_{tabel}$. Sebaliknya jika $r_{xy} \leq r_{tabel}$ maka soal dapat dianggap tidak valid. Terdapat 24 soal yang tidak valid (80%) dan sisanya (20 %) soal termasuk valid.

Reliabilitas adalah tingkat kekonsistenan dari suatu instrumen. Reliabilitas tes menyatakan bahwa suatu tes sudah teliti dan dapat dipercaya sesuai dengan kriteria yang telah ditentukan. Apabila suatu tes selalu memberikan hasil yang sama meskipun diteskan pada kelompok yang sama dan pada waktu atau kesempatan yang berbeda maka tes tersebut dikatakan reliabel ([Oliva & Blanco, 2023](#)). Analisis reliabilitas dilakukan terhadap sejumlah butir soal tes secara keseluruhan, namun analisis validitas dilakukan pada masing-masing butir soal tes dimana analisis menunjukkan bahwa soal PTS Biologi Kelas X adalah tidak reliabel.

Selain menggunakan metode CTT, soal PTS Biologi kelas X juga dianalisis dengan metode IRT model *Rasch* dengan menggunakan *Winstep*. Analisis menggunakan IRT dilakukan karena memiliki beberapa keunggulan baik dari aspek pengukuran, ketelitian, kemampuan membedakan dan fleksibilitas ([Zhong et al., 2023](#)). Analisis IRT dianggap lebih baik daripada CTT dalam memberikan informasi yang lebih akurat, teliti, dan mendalam tentang kemampuan individu dalam menghadapi tes atau instrumen penilaian. Untuk mengetahui hasil pengukuran menggunakan metode IRT dapat dilihat pada grafik *Wright map* soal PTS Biologi pada **Gambar 1**.



Gambar 1. Wright Map

Wright Map dapat menunjukkan bobot soal yang paling tinggi maupun rendah, juga siswa yang berkemampuan tinggi dan rendah. Hal ini dapat dilihat berdasarkan mistar *logit*nya. Jika soal semakin ke atas berarti bobot soal tersebut semakin besar yang berarti merupakan soal sulit. Semakin ke bawah posisinya berarti semakin rendah bobot soal tersebut. Artinya soal tersebut termasuk soal yang mudah. Begitu juga dengan posisi person (siswa), siswa yang berada pada posisi atas berarti dikatakan siswa tersebut memiliki *ability* tinggi. Jadi siswa dianggap siswa yang paling pandai. Sedangkan semakin ke bawah berarti siswa yang memiliki kemampuan rendah (Prieto et al., 2003).

Wright Map menunjukan bahwa soal nomor 25 memiliki bobot yang paling tinggi karena hanya dua siswa yang mampu menjawab soal tersebut. Siswa yang memiliki kemampuan paling tinggi yaitu siswa dengan kode L07 dan P08 (artinya siswa nomor 7 berjenis kelamin laki-laki dan nomor 8 berjenis kelamin perempuan) mampu menjawab soal nomor 25 serta soal lainnya yang memiliki nilai logit dibawahnya. Soal yang paling baik adalah soal yang berada pada posisi mistar logit M. Pada instrumen soal PTS Biologi Kelas X ini adalah soal nomor 2, 4, 8, 14 dan 21. *Wright Map* juga menunjukkan bahwa banyak siswa yang memiliki nilai logit di bawah nilai logit soal. Hal ini berarti siswa-siswa tersebut memiliki kemampuan yang lebih rendah daripada bobot soal.

Item STATISTICS: MEASURE ORDER													
ENTRY	TOTAL	TOTAL	MODEL	INFIT	OUTFIT	PTMEASUR-AL	EXACT	MATCH					
NUMBER	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	Item
25	9	30	1.25	.41	.91	-.50	.88	-.61	.42	.19	73.3	70.2	S25
20	11	30	.94	.39	.91	-.68	.90	-.77	.41	.20	66.7	64.8	S20
3	12	30	.80	.38	1.03	.31	1.03	.26	.14	.20	60.0	62.3	S3
19	13	30	.65	.38	1.07	.81	1.07	.79	.04	.20	50.0	60.3	S19
1	14	30	.51	.37	.95	-.69	.95	-.60	.32	.20	73.3	59.0	S1
7	14	30	.51	.37	.82	-2.42	.81	-2.41	.62	.20	80.0	59.0	S7
2	15	30	.37	.37	.95	-.63	.95	-.64	.31	.20	66.7	58.2	S2
14	16	30	.24	.37	1.18	2.34	1.22	2.47	-.25	.20	46.7	58.4	S14
4	17	30	.10	.38	1.12	1.44	1.15	1.50	-.10	.20	53.3	59.1	S4
8	17	30	.10	.38	1.02	.30	1.03	.38	.14	.20	60.0	59.1	S8
21	17	30	.10	.38	1.00	.04	.98	-.14	.20	.20	46.7	59.1	S21
16	18	30	-.05	.38	1.10	1.07	1.15	1.31	-.08	.19	63.3	60.7	S16
22	18	30	-.05	.38	.94	-.65	.92	-.68	.35	.19	63.3	60.7	S22
28	18	30	-.05	.38	.92	-.86	.89	-.93	.40	.19	63.3	60.7	S28
30	18	30	-.05	.38	.89	-1.19	.86	-1.26	.47	.19	63.3	60.7	S30
12	19	30	-.19	.39	1.02	.22	.99	-.02	.16	.19	60.0	63.5	S12
13	19	30	-.19	.39	.91	-.78	.88	-.86	.41	.19	66.7	63.5	S13
18	19	30	-.19	.39	1.06	.52	1.09	.70	.03	.19	60.0	63.5	S18
26	19	30	-.19	.39	1.04	.41	1.12	.89	.04	.19	60.0	63.5	S26
27	19	30	-.19	.39	1.01	.13	.98	-.07	.18	.19	60.0	63.5	S27
6	20	30	-.34	.39	1.07	.57	1.16	1.00	-.04	.18	66.7	66.7	S6
10	20	30	-.34	.39	.95	-.29	.91	-.50	.31	.18	66.7	66.7	S10
11	20	30	-.34	.39	.99	.00	.96	-.19	.22	.18	66.7	66.7	S11
17	20	30	-.34	.39	1.08	.62	1.10	.66	-.02	.18	66.7	66.7	S17
23	20	30	-.34	.39	1.02	.19	1.01	.14	.14	.18	66.7	66.7	S23
24	20	30	-.34	.39	1.03	.28	1.02	.16	.11	.18	66.7	66.7	S24
5	21	30	-.50	.40	.95	-.24	.92	-.37	.30	.17	70.0	70.0	S5
9	21	30	-.50	.40	.99	-.01	1.00	.05	.19	.17	70.0	70.0	S9
15	22	30	-.67	.42	1.00	.04	.97	-.04	.19	.17	73.3	73.3	S15
29	22	30	-.67	.42	1.03	.21	1.06	.35	.08	.17	73.3	73.3	S29
MEAN	17.6	30.0	.00	.39	1.00	.0	1.00	.0			64.1	63.9	
P.SD	3.2	.0	.47	.01	.08	.8	.10	.9			7.6	4.4	

Gambar 2. Tingkat Kesukaran dengan Model Rasch

Hasil analisis *Output Measure Order* pada **Gambar 2** menunjukkan bahwa soal dengan bobot tertinggi yaitu soal nomor 25 memiliki nilai *measure* 1,25 dan soal nomor 15 dan 19 dengan bobot soal lebih rendah memiliki nilai *measure* -0,67. Soal yang mampu dijawab oleh siswa yang dianggap memiliki ability tinggi adalah soal nomor 20 dengan nilai *measure* 0,94. Dilihat model S.E (standar error), soal yang memiliki nilai S.E lebih dari 0,5 atau mendekati 0,5 maka dikatakan sebagai soal yang bagus karena memiliki soal tersebut memiliki ketelitian yang baik. Pada tabel tersebut bisa dikatakan bahwa soal nomor 1, 2, 3, 7, 14, dan 19 tergolong soal yang bagus.

Tingkat kesukaran setiap butir soal dapat dianalisis dengan melihat *measure logit*. Tingkat kesukaran butir soal dikategorikan sulit, sedang dan mudah berdasarkan nilai standar deviasi (SD) dan mean *measure*. Standar deviasi dari *mean measure* adalah 0,47 sehingga tingkat kesukaran butir soal PTS Biologi Kelas X dapat dikelompokkan pada **Tabel 3**.

Tabel 3. Ringkasan Tingkat Kesukaran Butir Soal PTS Biologi dengan Model *Rasch*

Nilai Measure Logit	Kategori	Nomor Soal	Jumlah Butir Soal	Persentase (%)
Measure logit > 0,47	Sulit	1, 3, 7, 19, 20, 25	6	20
-0,47 < Measure Logit < 0,47	Sedang	2, 4, 6, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 21, 22, 23, 24, 26, 27, 28, 30	20	66,7
Measure Logit < -0,47	Mudah	5, 9, 15, 29	4	13,3

Perbandingan persentase bobot soal berdasarkan SD ini belum mendekati kriteria perbandingan yang baik menurut Sudjana (2014) yaitu proporsi perbandingan soal mudah, sedang dan sulit adalah 3-5-2 dimana soal dengan kategori sulit sebanyak 20%, sedang sebanyak 66,7% dan mudah sebanyak 13,3%. Hal ini karena adanya perbedaan kriteria penentuan batas suatu soal dikategorikan sulit, sedang dan mudah. Pada CTT, kategori soal sedang adalah pada rentang antara 0,3 sampai 0,7 sedangkan pada *Rasch* model adalah berdasarkan SD yaitu soal kategori sedang pada rentang antara -0,47 sampai 0,47. Hasil hitungan indeks sparasi ditunjukkan pada **Gambar 3**.

SUMMARY OF 30 MEASURED Item								
	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFI MNSQ	INFI ZSTD	OUTFI MNSQ	OUTFI ZSTD
MEAN	17.6	30.0	.00	.39	1.00	.02	1.00	.02
SEM	.6	.0	.09	.00	.01	.16	.02	.17
P.SD	3.2	.0	.47	.01	.08	.85	.10	.92
S.SD	3.3	.0	.48	.01	.08	.86	.10	.94
MAX.	22.0	30.0	1.25	.42	1.18	2.34	1.22	2.47
MIN.	9.0	30.0	-.67	.37	.82	-2.42	.81	-2.41
REAL RMSE	.39	TRUE SD	.25	SEPARATION	.65	Item	RELIABILITY	.30
MODEL RMSE	.39	TRUE SD	.26	SEPARATION	.68	Item	RELIABILITY	.32
S.E. OF Item MEAN = .09								
Item RAW SCORE-TO-MEASURE CORRELATION = -1.00								
Global statistics: please see Table 44.								
UMEAN=.0000 USCALE=1.0000								

Gambar 3. Separation Item

Asumsi pada pemodelan *Rasch* bahwa daya pembeda pada seluruh butir soal memiliki taraf yang sama. Pengelompokan butir soal menggunakan **Rumus 1** (persamaan strata (H)).

$$H = \frac{[(4 \times SEPARATION) + 1]}{3} \quad (1)$$

Hasil menunjukan bahwa nilai item separation adalah 0,65. Oleh karena itu dapat dihitung bahwa nilai H adalah 1,2 (dibulatkan menjadi 1). Hal ini berarti butir soal tidak dapat dikelompokkan menjadi tiga (rendah, sedang, tinggi) melainkan hanya dapat digolongkan menjadi soal yang rendah saja.

Item CATEGORY/OPTION/DISTRACTOR FREQUENCIES: MEASURE ORDER											
ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	ABILITY MEAN	P.SD	S.E. MEAN	INFT MNSQ	OUTF MNSQ	PTMA CORR.	Item
25	D	0	8	27	.14	.30	.11	.8	.8	-.35	S25
	C	0	13	43	.34	.36	.10	1.0	1.0	-.08	
	A	1	9	30	.65	.43	.15	.9	.9	.42	
20	A	0	3	10	.04	.07	.05	.7	.7	-.27	S20
	D	0	4	13	.28	.32	.19	.9	.9	-.09	
	C	0	12	40	.29	.41	.12	1.0	1.0	-.17	
3	A	0	4	13	.21	.33	.19	.9	.9	-.16	S3
	C	0	9	30	.27	.50	.18	1.0	1.0	-.17	
	B	0	3	10	.47	.19	.13	1.1	1.1	.07	
19	D	0	2	7	.65	.23	.23	1.3	1.3	.17	
	E	1	12	40	.45*	.39	.12	1.0	1.0	.14	
	B	0	4	13	.10	.25	.15	.8	.8	-.26	S19
1	C	0	6	20	.41	.33	.15	1.1	1.1	.03	
	A	0	7	23	.48	.34	.24	1.2	1.2	.08	
	D	0	4	13	.48	.49	.28	1.2	1.3	.10	
7	E	0	1	3	.00	.00		.7	.7	-.17	S7
	A	0	3	10	.24	.34	.24	.9	.9	-.11	
	B	0	1	3	.28	.00		.9	.9	-.05	
2	C	0	11	37	.28	.18	.06	.9	.9	-.18	
	D	1	14	47	.52	.53	.15	1.0	1.0	.32	
	E	0	1	3	.00	.00		.7	.7	-.17	S2
14	C	0	3	10	.00	.00	.00	.7	.7	-.31	
	D	0	9	30	.17	.30	.10	.9	.8	-.33	
	A	0	3	10	.23	.13	.10	.9	.9	-.12	
4	B	1	14	47	.66	.40	.11	.8	.8	.62	
	D	0	2	7	.07	.35	.35	.8	.8	-.20	S4
	E	0	3	10	.23	.07	.05	.9	.9	-.12	
20	C	0	7	23	.27	.38	.15	1.0	1.0	-.15	
	A	0	3	10	.35	.50	.35	1.1	1.1	-.02	
	B	1	15	50	.51	.42	.11	.9	.9	.31	
14	A	0	2	7	-.08	.21	.21	.7	.7	-.29	S14
	C	0	9	30	.52	.52	.18	1.3	1.3	.22	
	D	0	1	3	.72	.00		1.5	1.4	.15	
4	E	0	2	7	.81	.24	.24	1.6	1.6	.28	
	B	1	16	53	.28*	.29	.07	1.1	1.1	-.25	
	C	0	7	23	.28	.31	.13	1.0	1.0	-.13	S4
8	D	0	3	10	.58	.32	.23	1.3	1.3	.16	
	B	0	3	10	.62	.61	.43	1.4	1.6	.19	
	A	1	17	57	.34*	.40	.10	1.1	1.1	-.10	
21	A	0	3	10	-.10	.07	.05	.7	.6	-.38	S21
	B	0	1	3	.28	.00		.9	.9	-.05	
	D	0	1	3	.42	.00		1.1	1.1	.02	
16	C	0	8	27	.46	.48	.18	1.2	1.2	.12	
	E	1	17	57	.43*	.39	.10	1.0	1.0	.14	
	D	0	3	10	.04	.18	.12	.8	.7	-.27	S16
22	C	0	3	10	.14	.35	.25	.9	.8	-.19	
	B	0	6	20	.40	.18	.08	1.1	1.1	.02	
	E	0	1	3	.72	.00		1.5	1.4	.15	
12	A	1	17	57	.45*	.48	.12	1.0	1.0	.20	
	E	0	1	3	-.14	.00		.6	.6	-.23	S12
	A	0	4	13	.25	.30	.17	1.0	.9	-.12	
28	D	0	7	23	.60	.50	.20	1.4	1.4	.29	
	C	1	18	60	.35*	.37	.09	1.1	1.1	-.08	
	A	0	10	33	.14	.28	.09	.9	.8	-.41	S28
30	C	0	2	7	.51	.37	.37	1.2	1.2	.08	
	D	1	18	60	.50*	.43	.10	.9	.9	.35	
	E	0	2	7	.00	.00	.00	.7	.7	-.25	S30
17	D	0	6	20	.12	.35	.15	.9	.8	-.31	
	A	0	3	10	.28	.12	.08	1.0	.9	-.08	
	B	0	1	3	.57	.00		1.3	1.2	.08	
24	C	1	18	60	.52*	.43	.10	.9	.9	.40	
	B	0	1	3	.00	.00		.7	.7	-.17	S24
	A	0	1	3	.14	.00		.8	.8	-.11	
5	C	0	5	17	.26	.35	.18	1.0	1.0	-.13	
	D	0	1	3	.57	.00		1.4	1.3	.08	
	E	1	18	60	.54	.43	.10	.9	.9	.47	
11	D	0	1	3	.00	.00		.7	.7	-.17	S11
	E	0	2	7	.07	.07	.07	.8	.8	-.20	
	C	0	6	20	.33	.27	.12	1.1	1.0	-.06	
13	A	0	1	3	.42	.00		1.1	1.1	.02	
	B	1	20	67	.44	.47	.11	1.0	1.0	.22	
	D	0	3	10	.33	.30	.21	1.1	1.0	-.04	S13
18	C	0	6	20	.44	.39	.18	1.2	1.2	.07	
	A	1	20	67	.37*	.45	.10	1.1	1.1	-.02	
	E	0	7	23	.18	.36	.15	.9	.9	-.26	S18
26	D	0	2	7	.50	.22	.22	1.2	1.2	.08	
	A	0	1	3	.72	.00		1.5	1.5	.15	
	C	1	20	67	.42*	.44	.10	1.0	1.0	.14	
27	E	0	2	7	.07	.21	.21	.8	.8	-.20	S27
	B	0	2	7	.35	.07	.07	1.1	1.0	-.02	
	A	0	2	7	.36	.36	.36	1.1	1.1	-.01	
6	D	0	4	13	.39	.33	.19	1.1	1.1	.01	
	C	1	20	67	.41	.46	.11	1.0	1.0	.11	
	E	0	2	7	.00	.00	.00	.7	.7	-.25	S6
10	A	0	3	10	.05	.37	.26	.8	.8	-.27	
	D	0	2	7	.35	.07	.07	1.1	1.0	-.02	
	C	0	2	7	.43	.29	.29	1.2	1.1	.03	
9	A	1	21	70	.46	.43	.10	1.0	1.0	.30	
	D	0	3	10	.04	.29	.21	.8	.8	-.27	S9
	A	0	3	10	.14	.20	.14	.9	.8	-.19	
15	B	0	3	10	.59	.44	.31	1.4	1.4	.17	
	C	1	21	70	.43*	.41	.09	1.0	1.0	.19	
	D	0	1	3	.00	.00		.7	.7	-.17	S15
29	B	0	1	3	.14	.00		.9	.8	-.11	
	E	0	5	17	.26	.35	.18	1.0	1.0	-.13	
	C	0	1	3	.57	.00		1.4	1.3	.08	
2	A	1	22	73	.43*	.44	.10	1.0	1.0	.19	
	D	0	1	3	-.14	.00		.7	.6	-.23	S2
	B	0	2	7	.36	.36	.36	1.1	1.1	-.01	
21	E	0	4	13	.37	.48	.28	1.1	1.2	-.01	
	C	0	1	3	.57	.00		1.4	1.3	.08	
	A	1	22	73	.40*	.41	.09	1.0	1.0	.08	

Gambar 4. Hasil Analisis Keefektifan Distraktor

Distractor frequencies menunjukkan keefektifan distraktor. Pengecoh (distraktor) dikatakan berfungsi jika ada siswa yang memilih opsi tersebut. Berdasarkan tabel tersebut terlihat bahwa pada Soal Nomor 3 (S3) dengan kunci jawaban E dipilih 12 orang siswa dari 30 siswa yang ikut mengerjakan ujian atau 40% siswa yang menjawab benar. Pilihan jawaban A, B, C dan D yang berfungsi sebagai pengecoh juga dipilih oleh siswa. Hal ini berarti pengecoh tersebut dapat berfungsi. Opsi A dipilih sebanyak empat siswa (13%), opsi B dipilih tiga siswa (10%), opsi C dipilih oleh sembilan siswa (30%) dan opsi D dipilih oleh dua siswa (7%).

Keefektifan distraktor juga dilihat berdasarkan *average ability* setiap opsi jawaban pada butir soal tersebut. Jika terdapat peningkatan *average ability* berarti distraktor berfungsi (Morville et al., 2024). Misalnya pada S3, *average ability* opsi D adalah 0,07, opsi E adalah -0,23, opsi C adalah 0,27,

opsi A adalah 0,35 dan opsi B (kunci jawaban) adalah 0,51. Berturut-turut terdapat peningkatan *average ability*, berarti keefektifan distraktor berfungsi. Pada S14, *average ability* opsi jawaban adalah -0,08; 0,52; 0,72; 0,81 dan 0,28. Ringkasan hasil analisis keefektifan distraktor disajikan pada **Tabel 4**.

Tabel 4. Ringkasan Keefektifan Distraktor Butir Soal PTS Biologi Kelas X dengan Model *Rasch*

Efektifitas Distraktor	Nomor Soal	Jumlah Butir Soal	Persentase (%)
Berfungsi	1, 2, 5, 7, 8, 10, 11, 12, 13, 20, 24, 25, 30	13	43,3
Tidak Berfungsi	3, 4, 6, 9, 14, 15, 16, 17, 18, 19, 21, 22, 23, 26, 27, 28, 29	17	56,7

Item STATISTICS: MISFIT ORDER													
ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFINIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PTMEASUR-AL CORR.	EXP.	EXACT OBS%	MATCH EXP%	Item
14	16	30	.24	.37	1.18	2.34	1.22	2.47	A-.25	.20	46.7	58.4	S14
6	20	30	-.34	.39	1.07	.57	1.16	1.00	B-.04	.18	66.7	66.7	S6
4	17	30	.10	.38	1.12	1.44	1.15	1.50	C-.10	.20	53.3	59.1	S4
16	18	30	-.05	.38	1.10	1.07	1.15	1.31	D-.08	.19	63.3	60.7	S16
26	19	30	-.19	.39	1.04	.41	1.12	.89	E .04	.19	60.0	63.5	S26
17	20	30	-.34	.39	1.08	.62	1.10	.66	F-.02	.18	66.7	66.7	S17
18	19	30	-.19	.39	1.06	.52	1.09	.70	G .03	.19	60.0	63.5	S18
19	13	30	.65	.38	1.07	.81	1.07	.79	H .04	.20	50.0	60.3	S19
29	22	30	-.67	.42	1.03	.21	1.06	.35	I .08	.17	73.3	73.3	S29
3	12	30	.80	.38	1.03	.31	1.03	.26	J .14	.20	60.0	62.3	S3
8	17	30	.10	.38	1.02	.30	1.03	.38	K .14	.20	60.0	59.1	S8
24	20	30	-.34	.39	1.03	.28	1.02	.16	L .11	.18	66.7	66.7	S24
12	19	30	-.19	.39	1.02	.22	.99	-.02	M .16	.19	60.0	63.5	S12
23	20	30	-.34	.39	1.02	.19	1.01	.14	N .14	.18	66.7	66.7	S23
27	19	30	-.19	.39	1.01	.13	.98	-.07	O .18	.19	60.0	63.5	S27
9	21	30	-.50	.40	.99	-.01	1.00	.05	o .19	.17	70.0	70.0	S9
15	22	30	-.67	.42	1.00	.04	.97	-.04	n .19	.17	73.3	73.3	S15
21	17	30	.10	.38	1.00	.04	.98	-.14	m .20	.20	46.7	59.1	S21
11	20	30	-.34	.39	.99	.00	.96	-.19	l .22	.18	66.7	66.7	S11
1	14	30	.51	.37	.95	-.69	.95	-.60	k .32	.20	73.3	59.0	S1
2	15	30	.37	.37	.95	-.63	.95	-.64	j .31	.20	66.7	58.2	S2
5	21	30	-.50	.40	.95	-.24	.92	-.37	i .30	.17	70.0	70.0	S5
10	20	30	-.34	.39	.95	-.29	.91	-.50	h .31	.18	66.7	66.7	S10
22	18	30	-.05	.38	.94	-.65	.92	-.68	g .35	.19	63.3	60.7	S22
28	18	30	-.05	.38	.92	-.86	.89	-.93	f .40	.19	63.3	60.7	S28
13	19	30	-.19	.39	.91	-.78	.88	-.86	e .41	.19	66.7	63.5	S13
20	11	30	.94	.39	.91	-.68	.90	-.77	d .41	.20	66.7	64.8	S20
25	9	30	1.25	.41	.91	-.50	.88	-.61	c .42	.19	73.3	70.2	S25
30	18	30	-.05	.38	.89	-1.19	.86	-1.26	b .47	.19	63.3	60.7	S30
7	14	30	.51	.37	.82	-2.42	.81	-2.41	a .62	.20	80.0	59.0	S7
MEAN	17.6	30.0	.00	.39	1.00	.0	1.00	.0			64.1	63.9	
P.SD	3.2	.0	.47	.01	.08	.8	.10	.9			7.6	4.4	

Gambar 5. Validitas Butir Soal

Analisis butir soal dengan model *Rasch* dapat dilihat dengan tiga kriteria yaitu nilai *outfit MNSQ*, *outfit ZSTD* dan *Point Measure Correlation (Pt-measure Corr)*. Suatu butir soal dikatakan valid jika $0,5 < outfit\ MNSQ < 1,5$; $-2,0 < outfit\ ZSTD < +2,0$ dan $0,4 < Point\ Measure\ Correlation\ (Pt-measure\ Corr) < 0,85$ (Ramadhani et al., 2024). Hasil analisis validitas disajikan pada **Gambar 5** dan **Tabel 5**.

Tabel 5. Ringkasan Validitas Butir Soal PTS Biologi Kelas X dengan Model *Rasch*

Validitas	Nomor Soal	Jumlah Butir Soal	Persentase (%)
Valid	3, 4, 6, 7, 8, 9, 13, 16, 17, 18, 19, 20, 23, 24, 25, 26, 28, 29, 30	19	63,3
Tidak Valid	1, 2, 5, 10, 11, 12, 14, 15, 21, 22 27	11	36,7

Analisis menunjukkan bahwa 19 butir soal (63,3%) valid dan 11 butir soal lainnya (36,7%) tidak valid. Jika dibandingkan dengan hasil analisis dengan menggunakan CTT, hasil ini menunjukkan adanya perbedaan. Hal ini karena kriteria kevalidan butir soal dengan menggunakan model *Rasch* lebih detail yaitu meliputi tiga kriteria (*outfit MNSQ*, *outfit ZSTD* dan *Point Measure Correlation (Pt-measure Corr)*) sedangkan pada CTT hanya berdasarkan nilai *r* hitung berdasarkan rumus korelasi Pearson (product Moment). Selain itu, hasil ringkasan validitas dengan model *Rasch* menerapkan aturan jika ada salah satu kriteria yang tidak terpenuhi maka soal dianggap tidak valid meskipun dua kriteria lainnya terpenuhi (Hizqiyah et al., 2023).

SUMMARY OF 30 MEASURED Person									
	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		
					MNSQ	ZSTD	MNSQ	ZSTD	
MEAN	17.6	30.0	.38	.39	1.00	-.05	1.00	-.06	
SEM	.5	.0	.08	.00	.02	.20	.02	.21	
P.SD	2.7	.0	.42	.02	.11	1.09	.12	1.11	
S.SD	2.8	.0	.42	.02	.11	1.11	.12	1.13	
MAX.	24.0	30.0	1.45	.47	1.22	2.22	1.24	2.29	
MIN.	13.0	30.0	-.29	.37	.79	-1.96	.78	-1.96	

REAL RMSE	.40	TRUE SD	.13	SEPARATION	.32	Person RELIABILITY	.09		
MODEL RMSE	.39	TRUE SD	.15	SEPARATION	.39	Person RELIABILITY	.13		
S.E. OF Person MEAN = .08									

Person RAW SCORE-TO-MEASURE CORRELATION = 1.00									
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .08 SEM = 2.63									

Gambar 6. Reliabilitas Person

SUMMARY OF 30 MEASURED Item									
	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		
					MNSQ	ZSTD	MNSQ	ZSTD	
MEAN	17.6	30.0	.00	.39	1.00	.02	1.00	.02	
SEM	.6	.0	.09	.00	.01	.16	.02	.17	
P.SD	3.2	.0	.47	.01	.08	.85	.10	.92	
S.SD	3.3	.0	.48	.01	.08	.86	.10	.94	
MAX.	22.0	30.0	1.25	.42	1.18	2.34	1.22	2.47	
MIN.	9.0	30.0	-.67	.37	.82	-2.42	.81	-2.41	

REAL RMSE	.39	TRUE SD	.25	SEPARATION	.65	Item RELIABILITY	.30		
MODEL RMSE	.39	TRUE SD	.26	SEPARATION	.68	Item RELIABILITY	.32		
S.E. OF Item MEAN = .09									

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

Gambar 7. Reliabilitas Item

Penentuan reliabilitas item dan reliabilitas person dengan Model *Rasch* dikategorikan sebagai berikut: lemah ($< 0,67$); cukup ($0,67-0,80$); bagus ($0,81-0,90$); bagus sekali ($0,91-0,94$); dan istimewa ($>0,94$) ([Nejati & Jabbari, 2023](#)). Hasil analisis menunjukkan nilai reliabilitas person adalah 0,03 artinya konsistensi jawaban siswa termasuk kategori lemah dan mungkin dapat dikategorikan sangat lemah, sedangkan nilai reliabilitas item adalah 0,30 artinya reliabilitas soal termasuk kategori lemah pula. Ringkasan hasil analisis instrumen, khususnya reliabilitas person dan reliabilitas item disajikan pada **Gambar 6** dan **Gambar 7**.

Tabel 6. Perbandingan Hasil Analisis Butir Soal PTS Biologi Kelas X dengan CTT dan Model *Rasch* (IRT)

Indikator	CTT	IRT
Tingkat kesukaran	2 kategori: mudah (13,3%) dan sedang (86,7%)	3 kategori: mudah (20%); sedang (66,7%) dan sulit(13,3%)
Daya pembeda	4 kategori: rendah (53,3%); cukup (26,7%); tinggi (16,7%) dan sangat tinggi (3,3%)	1 kategori: rendah
Efektifitas Distraktor	Berfungsi (16,7%) dan tidak berfungsi(83,3%)	Berfungsi (43,3%) dan tidak berfungsi (56,7%)
Validitas Reliabilitas	Valid (20%) dan tidak valid (80%) Sangat rendah ($r_{11}=0,855$)	Valid (63,7%) dan tidak valid (36,3%) Reliabilitas item : 0,03 (lemah) Reliabilitas pearson : 0,30 (lemah)

Tabel 6 menunjukkan terdapat perbedaan hasil analisis yang mencolok antara CTT dan IRT. Pada indikator tingkat kesukaran CTT terdapat dua kategori sedangkan pada IRT tiga kategori. Pada indikator daya pembeda CTT menggolongkan menjadi 4 kategori sedangkan IRT 1 kategori saja. Pada indikator efektifitas distraktor dan validitas analisis CTT menunjukkan persentase soal yang berfungsi dan valid lebih kecil dibandingkan soal yang tidak berfungsi dan tidak valid, namun pada analisis IRT persentase terlihat lebih seimbang. Sedangkan pada indikator reliabilitas terdapat kemiripan dimana hasil analisis sama-sama menunjukkan data yang sifatnya tidak reliabel.

Dari berbagai penjelasan diatas dapat diuraikan bahwa analisis kualitas soal adalah bagian penting dari pengumpulan soal yang berkualitas tinggi untuk digunakan dalam tes. Tujuan utama analisis butir soal adalah untuk mengevaluasi kualitas soal dalam hal memberikan informasi tentang karakteristik unik dari setiap butir soal ([Tesio et al., 2024](#)).

Dari analisis tersebut ditemukan bahwa soal yang dibuat tidak memenuhi standar umum kualitas soal karena sebagian besar butir soal dinyatakan tidak valid dan reliabel sehingga dapat diberikan saran bahwa butir soal tersebut tidak layak digunakan untuk pengujian penilaian tengah semester (PTS) selanjutnya, sehingga diperlukan perbaikan secara menyeluruh instrumen soal disesuaikan dengan standar instrumen soal yang umum ([Egger, 2020](#)). Selain itu sebelum dan sesudah instrumen soal diujikan kepada siswa guru mata pelajaran harus melakukan analisis dan evaluasi terkait apakah soal yang diujikan sudah sesuai dengan standar penyusunan soal. Sehingga didapatkan hanya 20% soal yang dianggap valid, sementara 80% tidak valid, dan reliabilitas soal juga sangat rendah dengan nilai $r_{11} = 0,079$. Ini mengindikasikan bahwa sebagian besar soal tidak layak digunakan untuk pengujian penilaian tengah semester (PTS)

Berdasarkan uraian analisis diatas diketahui bahwa instrumen soal PTS yang diujikan belum memenuhi kaidah atau ketentuan yang sesuai sehingga perlu dilakukan perbaikan. Hal ini senada dengan yang dikemukakan oleh Safitri et al. (2022) bahwa hasil dari analisis butir-butir soal memungkinkan untuk menentukan kualitas soal: baik, cukup baik, atau tidak baik. Ketika kualitas soal diketahui, soal yang kurang baik dapat dilakukan perbaikan atau revisi untuk membantu meningkatkan kualitas soal yang akan digunakan dalam kegiatan penilaian dan evaluasi.

KESIMPULAN

Berdasarkan kajian diatas dapat disimpulkan bahwa penelitian bertujuan untuk mengidentifikasi soal-soal yang layak dan dapat digunakan dalam penilaian, sehingga siswa dapat dievaluasi dengan instrumen yang valid dan andal. Hasil analisis kualitas soal dengan menggunakan *classical test theory* (CTT) menjelaskan bahwa tingkat kesulitan soal terdapat dua kategori yaitu soal kategori mudah (13,3%) dan sedang (86,7%). Daya pembeda terdapat empat kategori yaitu rendah (53,3%), cukup (26,7%), tinggi (16,7%) dan sangat tinggi (3,3%). Efektifitas distraktor yang berfungsi sebanyak 16,7% sedangkan yang tidak berfungsi 83,3%. Soal yang tergolong valid terdapat 20% sedangkan yang tergolong tidak valid 80%. Secara keseluruhan berdasarkan analisis CTT, butir soal PTS Biologi Kelas X ini adalah tidak reliabel dengan nilai $r_{11} = 0,079$ (sangat rendah).

Hasil analisis butir soal dengan *item response theory* (IRT) menunjukkan bahwa tingkat kesukaran soal terdapat tiga kategori yaitu mudah (20%), sedang (66,7%) dan sulit (13,3%). Daya pembeda hanya terdapat satu kategori yaitu rendah. Efektifitas distraktor yang berfungsi sebanyak 43,3%, sedangkan yang tidak berfungsi 56,7%. Validitas butir soal yang tergolong valid 63,7% dan tergolong tidak valid 36,3. Reliabilitas item sebesar 0,03 dengan kategori lemah sedangkan reliabilitas person sebesar 0,30 sehingga dikategorikan lemah pula.

DAFTAR PUSTAKA

- Alfarisa, F., & Purnama, D. N. (2019). *Analisis Butir Soal Ulangan Akhir Semester Mata Pelajaran Ekonomi SMA Menggunakan RASCH Model*, 11(2), 366-374. <https://doi.org/10.23887/jjpe.v11i2.20878>
- Andayani, F., Tindangen, M., & Haryanto, Z. (2017). Analisis Permasalahan Guru Terkait Perencanaan dan Pelaksanaan Perangkat Pembelajaran Biologi Melalui Model Problem Based Learning dan Media Realita di SMA. *Jurnal Pendidikan: Teori, Penelitian, Dan Pengembangan*, 2(10), 1425-1429. <http://journal.um.ac.id/index.php/jptpp/>
- Egger, R. (2020). *A problem raising device. Was Lehrende bei einem Verzicht auf Vorlesungen verlieren würden.* https://doi.org/10.1007/978-3-658-29049-8_11
- Eliza, E., Djudin, T., & Oktavianty, E. (2022). Development of Critical Thinking Ability Test using the Rasch Model on Substance Pressure Materials. *Prisma Sains: Jurnal Pengkajian Ilmu Dan Pembelajaran Matematika Dan IPA IKIP Mataram*, 10(4), 842-855. <https://doi.org/10.33394/jps.v10i4.5987>
- Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). Analisis Kualitas Soal Kemampuan Membedakan Rangkaian Seri dan Pararel Melalui Teori Tes Klasik Dan Model Rasch . *Indonesian Journal of Educational Research and Review*, 3(1), 11-19. <https://doi.org/10.23887/ijerr.v3i1.24080>
- Fauziana, A., & Dessy Wulansari, A. (2021). Analisis Kualitas Butir Soal Ulangan Harian di Sekolah Dasar dengan Model Rasch . *Ibriez: Jurnal Kependidikan Dasar Islam Berbasis Sains*, 6, 10-19. <https://doi.org/10.21154/ibriez.v6i1.112>
- Handayani, S. (2022). Analisis uas biologi kelas x dengan teori tes klasik dan item response theory (Rasch model). *Bio-Pedagogi*, 11(2), 76-90. <https://doi.org/10.20961/bio-pedagogi.v11i2.63177>
- Hizqiyah, I. Y. N., Widodo, A., Sriyati, S., & Ahmad, A. (2023). Development of a Digital Problem Solving Skills Test Instrument: Model Rasch Analysis. *Jurnal Penelitian Pendidikan IPA*, 9(4), 1658-1663. <https://doi.org/10.29303/jppipa.v9i4.2671>

- Lestari, N. I., Razak, A., Lufri, L., & Zulyusri, Z. (2022). Analisis Pelaksanaan Evaluasi Untuk Meningkatkan Mutu Pembelajaran Biologi Di Sekolah. *Bioilmi: Jurnal Pendidikan*, 8(2), 114–127. <https://doi.org/10.19109/bioilmi.v8i2.11629>
- Maulida, S., Fatma, S. N., Fajriyah, D. T., Zahroturrosyidah, R., Hidayat, R. R., & Nahriyah, A. (2023). Analisis Hasil Evaluasi Melalui Pemberian Skor Tes Objektif dan Essay Serta Buku Catatan Lengkap Dan Tidak Lengkap. *Jurnal Ilmiah Keislaman*, 9(1), 2302–4801. <https://doi.org/10.21154/ibriez.v6i1.112>
- Misri, M. A., Saifuddin, S., Akbar, R. O., & Kamelia, N. R. (2021). Development and evaluation of a HOTS-based test for matrix topic: A classical test and item response theory. *Beta: Jurnal Tadris Matematika*, 14(2), 126–141. <https://doi.org/10.20414/betajtm.v14i2.383>
- Morville, A. Le, Wagman, P., & Håkansson, C. (2024). A Rasch analysis of the Danish version of the occupational balance questionnaire (OBQ11). *Scandinavian Journal of Occupational Therapy*, 31(1). <https://doi.org/10.1080/11038128.2024.2327356>
- Myrzagaliyeva, S., Izbassarova, R., Kuanyshbekova, A., & Ibadullayeva, S. (2024). Assessment of effectiveness of practice teaching biology methods in collective learning: the pilot study. *BIO Web of Conferences*, 100, 1–7. <https://doi.org/10.1051/bioconf/202410001026>
- Nejati, R., & Jabbari, M. (2023). Evaluating the Validity of Meara and Rogers' Vocabulary Aptitude Test: a Rasch Measurement Model Analysis. *Issues in Language Studies*, 12(2), 37–55. <https://doi.org/10.33736/ils.5290.2023>
- Oliva, J. M., & Blanco, Á. (2023). Rasch analysis and validity of the construct understanding of the nature of models in Spanish-speaking students. *European Journal of Science and Mathematics Education*, 11(2), 344–359. <https://doi.org/10.30935/scimath/12651>
- Prasetya, W. A., & Pratama, A. T. (2023). Item quality analysis using the Rasch model to measure critical thinking ability in the material of the human digestive system of Biology subject in high school. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 27(1), 76–91. <https://doi.org/10.21831/pep.v27i1.58873>
- Prieto, L., Alonso, J., & Lamarca, R. (2003). Classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health and Quality of Life Outcomes*, 1, 1–13. <https://doi.org/10.1186/1477-7525-1-27>
- Purwaningsih, E. (2022). Pemahaman Guru Tentang Analisis Kualitas Tes Dan Butir Soal. *Perspektif*, 1(3), 225–230. <https://doi.org/10.53947/perspekt.v1i3.98>
- Ramadhani, R., Syahputra, E., & Simamora, E. (2024). The construct validity of self-regulated learning questionnaire for senior high school students: a Rasch model analysis. *International Journal of Evaluation and Research in Education*, 13(3), 2032–2038. <https://doi.org/10.11591/ijere.v13i3.26816>
- Sari, V. N. I., Utomo, A. P. Y., & Sumarwati. (2022). Kualitas Soal Bahasa Indonesia di SMP Muhammadiyah 1 Pontianak: Analisis Butir Soal Vina. *Jurnal Pendidikan Bahasa Dan Sastra Indonesia*, 11(2), 112–119. <https://doi.org/10.15294/jpbsi.v11i2.58091>
- Skyttberg, N., Kottorp, A., & Alenius, L. S. (2023). Sound psychometric properties of a short new screening tool for patient safety climate: applying a Rasch model analysis. *BMC Health Services Research*, 23(1), 1–11. <https://doi.org/10.1186/s12913-023-09768-y>
- Stano, F., Pellicciari, L., Porta, F. La, Piscitelli, D., Angilecchia, D., Signorelli, M., Giovannico, G., Pournajaf, S., & Caselli, S. (2024). Rasch Analysis of the Forgotten Joint Score in Patients With

Total Hip Arthroplasty. *Journal of Rehabilitation Medicine*, 56(8), 1–9. <https://doi.org/10.2340/jrm.v56.15774>

Subali, B., Yogyakarta, U. N., Surakarta, U. M., Aminah, N. S., & Maret, U. S. (2021). The Comparison of Item Test Characteristics Viewed from Classic and Modern Test Theory. *International Journal of Instruction*, 14(1), 647–660. <https://doi.org/10.29333/iji.2021.14139a>

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. *Journal Disability and Rehabilitation*, 46(3), 604–617. <https://doi.org/10.1080/09638288.2023.2169772>

Zhong, J., Ma, H. Y., Wang, X. M., Huang, X. J., & Xu, M. Z. (2023). *Rasch* analysis of the Chinese version of the clinically useful depression outcome scale in patients with major depressive disorder. *BMC Psychology*, 11(1), 1–9. <https://doi.org/10.1186/s40359-023-01255-7>