Bio-Pedagogi: Jurnal Pembelajaran Biologi
http://jurnal.uns.ac.id/index.php/pdg/
biopedagogi@fkip.uns.ac.id
p-ISSN: 2252-6897
e-ISSN: 2715-176X

Analisis kualitas instrumen penilaian materi keanekaragaman hayati melalui tes klasik dan *Rasch* model

Nila Permatasari [a,1*]
[a] Program Studi Pendidikan Biologi, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Muhammadiyah Surakarta, Kabupaten Sukoharjo, Jawa Tengah, 57162, Indonesia.
[1] permatanila20@gmail.com*
**Corresponding author*

INFORMASI ARTIKEL

Lini Masa Artikel
Draft diterima : 2024-06-17
Revisi diterima : 2024-11-77
Diterbitkan : 2025-04-24

Keywords
Biodiversity learning;
Item quality analysis;
Learning assessment;

ABSTRAK

Evaluasi pembelajaran berfungsi sebagai indikator Penilaian merupakan kegiatan yang melibatkan interpretasi data pengukuran yang sesuai dengan kriteria atau aturan-aturan tertentu. Instrumen penilaian yang efektif harus memenuhi persyaratan yang meliputi validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas distraktor. Penelitian ini bertujuan untuk membandingkan hasil analisis instrumen penilaian antara teori tes klasik dan *Rasch* model pada assesment materi Keanekaragaman Hayati Penelitian dilakukan pada kelas VII di salah satu SMP di Sukoharjo. Penelitian ini berupa penelitian deskriptif kuantitatif dengan teknik sampling acak. Data dianalisis menggunakan Microsoft Excel (dan *Winsteps*. Hasil analisis validitas melalui teori tes klasik dikategorikan berkualitas cukup/sedang (5 valid) dari melalui analisis *Rasch* model diperoleh kualitas soal yang baik (6 valid). Analisis reliabilitas butir soal melalui teori tes klasik sebesar 0,458 (sedang) dan *Rasch* model diperoleh person bernilai lemah (0,34) dan butir soal bernilai bagus (0,86). Tingkat kesukaran, menurut teori tes klasik dan *Rasch* model terdistribusi dalam tiga kelompok. Daya pembeda tes teori klasik terbagi pada 5 soal bernilai cukup, 4 soal bernilai tinggi, dan 1 soal bernilai sangat tinggi. Lalu, pada *Rasch* model terdapat empat kelompok butir soal yang dapat diidentifikasi, sedangkan untuk person hanya terdapat satu kelompok. Untuk efektivitas distraktor pada tes teori klasik dan *Rasch* model sebagian besar sudah menunjukkan kualitas baik sedangkan sebagian kecil distraktor tidak bekerja. Pendekatan *Rasch* model dianggap lebih baik dikarenakan lebih objektif, mudah dalam menafsirkan hasil, fleksibel, dan kuat secara statis.

ABSTRACT

Learning evaluation functions as an indicator. Assessment is an activity that involves the interpretation of measurement data in accordance with certain criteria or rules. An effective assessment instrument must meet the requirements including validity, reliability, level of difficulty, discriminatory power, and distractor effectiveness. This study aims to compare the results of the analysis of assessment instruments between classical test theory and the Rasch model on the assessment of Biodiversity material. The study was conducted in class VII at one of the junior high schools in Sukoharjo. This study is a quantitative descriptive study with random sampling techniques. The data were analyzed using Microsoft Excel (and *Winsteps*. The results of the validity analysis through the classical test theory were categorized as sufficient/moderate quality (5 valid) from the Rasch model analysis obtained good question quality (6 valid). The reliability analysis of the questions through the classical test theory was 0.458 (moderate) and the Rasch model obtained a person with a weak value (0.34) and a question with a good value (0.86). The level of difficulty, according to the classical test theory and the Rasch model, was distributed into three groups. The discriminatory power of the classical theory test was divided into 5 questions with sufficient value, 4 questions with high value, and 1 question with very high value. Then, in the Rasch model there were four groups of questions that could be identified, while for the person there was only one group. For the effectiveness of the distractors in the classical theory test and the Rasch model, most of them had shown good quality while a small number of distractors did not work. The Rasch model approach is considered better because it is more objective, easy to interpret the results, flexible, and statically strong.

Cara Sitasi Artikel Ini (*APA Style*):
Permatasari, N. (2025). Analisis kualitas instrumen penilaian butir soal materi keanekaragaman hayati melalui tes klasik dan *Rasch* model. *Bio-Pedagogi. 14*(1), 10-26. https://dx.doi.org/10.20961/bio-pedagogi.v14i1.88470.

PENDAHULUAN

Pendidikan adalah kunci untuk pengembangan bangsa (Tamrin, 2019). Pentingnya peran pendidik dan proses belajar dalam menentukan kualitas pendidikan menjadi bagian fundamental. Oleh karena itu, pendidikan berkualitas tidak bersifat mandiri, namun tergantung pada proses evaluasi yang efektif dalam menyertai pelaksanaan kegiatan pengajaran dan pembelajaran. Evaluasi adalah elemen penting dalam proses pembelajaran, karena melibatkan pengumpulan data untuk menilai sejauh mana hasil pembelajaran yang dimaksud telah dicapai (Izza, 2020). Dengan begitu, guru berperan penting dalam sistem pendidikan, harus memiliki tidak hanya kemampuan untuk menyampaikan pengetahuan tetapi juga kemampuan untuk menilai efektivitas kegiatan belajar. Hasil dari proses evaluasi tersebut memberikan wawasan tentang keberhasilan upaya tersebut.

Evaluasi pembelajaran berfungsi sebagai indikator pencapaian tujuan pembelajaran (Arifin, 2022). Selain mengukur sejauh mana tujuan tercapai, evaluasi juga berfungsi sebagai faktor motivasi untuk pengambilan keputusan. Ini melibatkan proses mengukur dan menilai. Membandingkan objek menggunakan satu unit pengukuran dalam mengukur, sedangkan mengevaluasi berkaitan tentang suatu cara membandingkan dengan standar penilaian yang baik dan buruk dalam menilai (Jumini, 2023). Penilaian atau *assessment* adalah kegiatan yang melibatkan interpretasi data pengukuran atau hasil yang sesuai dengan kriteria atau aturan-aturan tertentu (Pramana, 2019). Berbeda dengan pengukuran yang bersifat kuantitatif, penilaian bersifat kualitatif.

Penilaian pada peserta didik sangat menguntungkan dalam mengukur sejauh mana atau bagaimana proses belajar bekerja secara efisien, serta berfungsi sebagai sumber motivasi bagi peserta didik dalam pembelajaran (Suardipa, 2023). Guru menggunakan penilaian sebagai evaluasi untuk mendapatkan pemahaman yang menyeluruh tentang proses keberhasilan peserta didik, efektivitas topik yang disajikan, dan strategi belajar yang digunakan (Habibi, 2021). Hasil evaluasi untuk sekolah dapat berfungsi sebagai dasar untuk pengambilan keputusan dan pembuatan kebijakan, dan dapat bertindak sebagai acuan untuk menilai kualitas lembaga pendidikan (Nasution, 2022). Kemampuan peserta didik biasanya ditunjukkan oleh tingkat pencapaian mereka. Jika peserta didik memenuhi atau melebihi kriteria skor minimum, mereka dianggap memiliki kemampuan yang baik dan begitu jika sebaliknya. Namun, menggunakan skor tidak dapat digunakan sebagai sarana untuk mengevaluasi kinerja atau prestasi peserta didik. Oleh karena itu, keputusan yang dibuat berdasarkan hasil skor peserta didik dianggap tidak mencukupi dan kurang tepat. Guru memiliki tanggung jawab sehubungan dengan penilaian kemajuan peserta didik. Guru memiliki kewajiban penilaian tambahan selain merancang dan menerapkan kegiatan belajar, hal ini adalah tanggung jawab utama sebagai pendidik (Guangul, 2020). Hasil evaluasi menunjukkan sejauh mana tingkat kualitas dari proses belajar telah dilakukan (Magdalena, 2023). Untuk mencapai hal ini, guru harus memiliki berbagai kemampuan dalam melakukan evaluasi pembelajaran, yang mencakup perencanaan, membangun alat evaluasi, menganalisis data dan menafsirkan hasil (Sylvia, 2019). Selain itu, guru harus memiliki kemampuan untuk mengembangkan topik yang terstruktur dengan baik dan terlibat dalam analisis kritis.

Instrumen penilaian yang efektif harus memenuhi persyaratan keandalan. Faktor-faktor yang dapat diandalkan meliputi validitas, reliabilitas, tingkat kesukaran, efektivitas distraktor, dan daya beda (Pratiwi, 2022). Untuk menilai keandalan dan kualitas instrumen dalam mengevaluasi suatu hasil belajar maka perlu dilakukan uji coba instrumen terlebih dahulu (Suratman, 2019). Pengujian internal dan eksternal dapat dilakukan pada instrumen. Pengujian internal dilakukan untuk mengevaluasi akurasi, struktur dan bahasa konten. Penilaian ahli diperlukan untuk menentukan apakah struktur instrumen sesuai dengan aturan yang digunakan dalam kompilasi instrumen. Untuk memvalidasi alat evaluasi, penting untuk melakukan tes eksternal, yaitu uji lapangan. Pengujian lapangan dapat dilakukan pada subjek yang serupa dengan yang sedang dievaluasi. Kemudian dilakukan analisis hasil uji coba, meliputi analisis validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas distraktor.

Validitas mengacu pada tingkat akurasi dan presisi yang ditunjukkan oleh alat ukur dalam memenuhi tujuan evaluasinya, yaitu dalam mengukur konstruksi yang dimaksudkan (Fernando, 2023). Reliabilitas ditentukan oleh sejauh mana hasil evaluasi dapat dipercaya yang ditandai dengan pengukuran pada subjek yang sama menghasilkan hasil penilaian yang konsisten (Ida, 2021). Tingkat kesukaran dalam evaluasi, mencerminkan keakuratan individu dalam memberikan jawaban dari butir soal yang diberikan. Efektivitas distraktor mengacu pada sejauh mana distraktor tersebut dapat mengecoh peserta didik yang tidak mengetahui jawaban yang benar. Soal atau item yang baik ditandai dengan tingkat kesulitan yang tidak terlalu menantang atau terlalu sederhana (Jumini, 2023). Daya beda soal mencerminkan tingkat kemampuan dan keterampilan dalam berpikir di antara peserta didik, hal ini berkaitan menentukan kemampuan peserta didik dalam berpikir kritis (Puspita, 2021). Untuk menentukan validitas, reliabilitas, tingkat kesukaran, daya pembeda dan efektivitas distraktor, penyelidikan menyeluruh dari evaluasi hasil belajar diperlukan (Tarmizi, 2020). Hal ini dapat dicapai dengan menggunakan pendekatan teori tes klasik *(Classical Test Theory (CTT))* atau menggunakan *Rasch* Model *(Item Response Theory (IRT))*.

Teori Tes Klasik (CTT), juga dikenal sebagai teori pengukuran klasik, adalah kerangka kerja statistik yang digunakan dalam psikometri untuk merancang, mengembangkan, dan menafsirkan tes (Purwaningsh, 2022). CTT menjelaskan korelasi antara hasil tes dan skor nyata yang tidak diamati. Skor mewakili kemampuan atau karakteristik peserta didik yang diamati secara langsung dan dinilai melalui tes. Kesalahan pengukuran merupakan faktor-faktor yang menyebabkan skor tes individu menyimpang dari skor sebenarnya. Kesalahan pengukuran dapat timbul dari berbagai sumber, termasuk kelelahan, kecemasan, kondisi lingkungan, dan variabel lainnya. Teori tes klasik didasarkan pada skor yang diamati, yang merupakan kombinasi dari skor aktual dan skor kesalahan pengukuran (Antara, 2019). Teori tes klasik menawarkan beberapa manfaat, termasuk kesederhanaan dan kemudahan pemahaman. Tes ini menyediakan berbagai teknik validasi dan standar untuk menghitung reliabilitas dan validitas tes. Selain itu, ia memiliki aplikasi yang luas di bidang pengukuran. Namun, Teori Tes Klasik (CTT) memiliki keterbatasan karena asumsi kaku, termasuk persyaratan homogenitas dan normalitas. Tes ini tidak mempertimbangkan karakteristik individu dari peserta didik, seperti gaya belajar dan strategi menjawab soal, dan tidak dapat tidak dapat memberikan informasi tentang performa individu pada butir soal tertentu.

Rasch model merupakan cara modern yang awalnya dikembangkan oleh Dr. Georg *Rasch*, seorang matematikawan Denmark, dengan tujuan mengatasi keterbatasan metode teori tes klasik. *Rasch* model membangun skala pengukuran dengan interval yang konsisten dengan menggunakan skor mentah dalam berbagai cara, sehingga menghasilkan data yang akurat tentang subjek uji dan kualitas subjek (Pudjiati, 2023). *Rasch* model adalah teknik psikometrik yang digunakan untuk memeriksa data kategoris, seperti tanggapan terhadap penilaian pilihan ganda atau survei (kuesioner). Sederhananya, cara ini berfokus pada hubungan antara kemampuan peserta didik (tingkat penguasaan) dan kesulitan item tes (Laliyo, 2021). Ini berbeda dengan Teori Tes Klasik (CTT) yang lebih mendasarkan pada skor total. *Rasch* model menawarkan beberapa keunggulan atas teori tes klasik (Polat, 2022). *Rasch* model memungkinkan untuk analisis yang lebih komprehensif dari data tes, yang membolehkan untuk mengidentifikasi item soal yang bermasalah yang mungkin terlalu mudah atau terlalu sukar untuk ujian. Selain itu, ini memfasilitasi penciptaan instrumen yang dapat diandalkan dan valid untuk penilaian pendidikan dan psikologis.

Dalam penelitian ini mencoba menjelaskan hasil analisis perbandingan kualitas instrumen penilaian dalam aspek validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan keefektifan distraktor lewat dua pendekatan seperti yang sudah dijabarkan tersebut, yakni cara klasik (teori tes klasik) dan modern (*Rasch* model). Instrumen tes yang dianalisis pada riset ini adalah instrumen penilaian untuk mengukur pemahaman materi Keanekaragaman Hayati di salah satu kelas VII di SMP Negeri daerah Sukoharjo. Dengan begitu, penelitian ini bertujuan untuk membandingkan hasil instrumen penilaian antara teori tes klasik dengan *Rasch* model pada soal pemahaman tentang materi Keanekaragaman Hayati pada salah satu kelas VII di SMP Negeri daerah Sukoharjo.

METODE

Penelitian ini merupakan penelitian deskriptif untuk menciptakan dan memvalidasi temuan (Bungin, 2021). Item soal tes yang diuji mencakup soal pilihan ganda dari materi keanekaragaman hayati, yang terdiri dari 10 butir soal dan masing-masingnya terdapat empat pilihan jawaban. Penelitian ini menjelaskan analisis dari hasil instrumen item soal dan person (khusus *Rasch* model) pada pemahaman materi keanekaragaman hayati. Analisis ini berfokus pada: validitas, reliabilitas, tingkat kesukaran, daya pembeda dan efektivitas distraktor. Sampel diambil dengan teknik sampling acak, di mana satu kelas dipilih sebagai subjek penelitian. Hal ini digunakan untuk memastikan sampel yang dipilih mewakili seluruh populasi sangat penting untuk secara akurat menggeneralisasi hasil penelitian (Firmansyah, 2022). Uji coba lapangan dilaksanakan pada 32 peserta didik (14 perempuan, 18 laki-laki). Data dianalisis dengan Microsoft Excel dan perangkat lunak *Winsteps*.

Analisis dengan teori tes klasik dilakukan dengan Microsoft Excel, uji validitas dilakukan untuk menghitung nilai koefisien korelasi. Sementara itu, uji reliabilitas dilakukan untuk menghitung nilai koefisien reliabilitas yang digunakan untuk menilai konsistensi dan stabilitas hasil pengukuran instrumen penelitian. Uji tingkat kesukaran (TK) dilakukan untuk menghitung nilai indeks TK yang digunakan untuk mengetahui tingkat kemudahan atau kesulitan butir soal dalam suatu tes. Uji daya pembeda (D) dilakukan untuk menghitung nilai indeks D yang digunakan untuk mengetahui kemampuan butir soal dalam memisahkan peserta didik yang memiliki kemampuan tinggi dan rendah. Uji efektivitas distraktor dilakukan untuk menghitung nilai indeks daya tarik (DI) dan daya tolak (DT) yang digunakan untuk mengetahui seberapa efektif distraktor dalam menarik dan menolak peserta didik yang tidak menguasai materi.

Rasch model memakai perangkat lunak *Winsteps*, uji validitas untuk menganalisis kualitas instrumen pengukuran. Uji reliabilitas dilakukan untuk menilai konsistensi dan stabilitas hasil pengukuran instrumen penelitian. Uji tingkat kesukaran dilakukan untuk menilai tingkat kesukaran butir-butir soal dalam suatu tes. Uji daya pembeda untuk meningkatkan kualitas tes dan memastikan bahwa tes tersebut dapat mengukur kemampuan peserta didik secara akurat. Uji efektivitas distraktor dilakukan untuk menilai kemampuan distraktor (pilihan jawaban yang salah) dalam suatu butir soal untuk menarik peserta didik yang tidak memiliki kemampuan yang sesuai untuk menjawab butir soal tersebut dengan benar.

HASIL DAN PEMBAHASAN

Validitas adalah tingkat dimana instrumen mengukur apa yang dimaksudkan untuk diukur. Validitas ini meliputi tiga jenis, yaitu validitas isi, konstruk, serta kriteria (Azwar, 2019). Validitas isi mengacu pada sejauh mana instrumen pengukuran secara akurat menggambarkan seluruh analisis domain yang sedang dievaluasi. Validitas konstruk adalah jawaban apakah semua butir menjadi sampel representatif dari semua butir yang nantinya dibuat, dan terkait dengan ekivalensi batas dari domain yang dapat diukur. Dengan kata lain, nilai validitas konstruk instrumen pengujian berarti sejauh mana ia menilai atribut atau kemampuan peserta didik. Validitas kriteria berkaitan dengan kemampuan untuk memprediksi kesuksesan masa depan peserta didik, dapat dipastikan melalui korelasi kriteria spesifik dengan instrumen. Validitas penilaian instrumen pada pendekatan teori tes klasik bisa ditentukan dengan memakai rumus korelasi moment produk ataupun metode korelasi point biserial (rpbi) (Bichi, 2019). Bila dinyatakan valid jika koefisien korelasi rxy > r tabel, serta bila rxy ≤ r maka soal bisa dinyatakan tidak valid. Item soal yang dianggap valid adalah soal yang mampu mengukur dengan tepat apa yang ingin diukur. Dengan kata lain, soal yang valid bebas dari kesalahan dan memberikan hasil yang akurat. Ketika memeriksa validitas suatu item soal menggunakan *Rasch* model, penting untuk menilai apakah itu memenuhi kriteria tertentu dalam hal kualitas, sebagai berikut (Erfan et al., 2020).

a). Skor *Outfit MNSQ (Mean Square)* yang diterima yakni: 0,5 < *Outfit – MNSQ* < 1,5
b). Skor *Outfit ZSTD (Z – Standard)* yang diterima yakni: –2,0 < *ZSTD* < +2,0
c). Skor *Pt Measure Corr (Point Measure Correlation):* 0,4 < *Point Measure Corr* < 0,85

Jika butir soal memenuhi sedikitnya dua kriteria maka dianggap valid; jika hanya memenuhi satu kriteria, butir soal dianggap dapat diperbaiki; serta jika tidak memenuhi salah satu dari tiga kriteria ini maka dibuang atau dihapus. Setelah melakukan analisis dari tes dengan *Rasch* model dan teori tes klasik, perbedaan hasil analisis validitas dari butir soal diperoleh dan terperinci dalam **Tabel 1**. Dalam analisis *Rasch* model ditampilkan pada **Gambar 1** (*Outfit MNSQ*, *Outfit ZSTD*, serta *Point Measure Correlation (Pt Measure Corr)*). Ketiga indikator ini memberikan gambaran yang lebih objektif mengenai sejauh mana butir soal berfungsi dengan baik dalam mengukur kemampuan siswa. Jika nilai *Outfit MNSQ* berada dalam rentang 0,5 hingga 1,5, nilai ZSTD berkisar antara -2 hingga +2, serta *Pt Measure Corr* positif (di atas 0,2), maka butir soal dinilai valid menurut *Rasch*. Temuan ini kemudian dibandingkan dengan hasil analisis validitas menggunakan teori tes klasik, yang umumnya didasarkan pada korelasi antara skor butir dan total skor. Perbandingan ini memberikan wawasan yang lebih menyeluruh tentang kualitas instrumen yang digunakan dalam penelitian.

Tabel 1. Perbandingan Hasil Analisis Validitas Butir Soal melalui Teori Tes Klasik dan *Rasch* Model

No	Hasil	No Soal	
		Teori Tes Klasik	***Rasch* Model**
1.	Valid	1,4,5,6,8	4, 5,6,7,8,9
2.	Tidak Valid	2,3,7,9,10	1,2,3,10

13-851WS - Notepad

File Edit Format View Help

TABLE 13.1 C:\Users\acer\Desktop\DATA 2.prn ZOU851WS.TXT Apr 14 2024 7:37
INPUT: 32 Person 10 Item REPORTED: 32 Person 10 Item 2 CATS WINSTEPS 4.5.2

Person: REAL SEP.: .72 REL.: .34 ... Item: REAL SEP.: 2.50 REL.: .86

Item STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	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
10	4	32	3.13	.57	1.09	.3	1.05	.33	.25	.32	87.5	87.4	S10
7	11	32	1.57	.42	1.00	.0	1.00	.10	.44	.44	68.8	72.9	S7
2	20	32	.12	.41	1.17	.9	1.15	.60	.33	.45	65.6	72.1	S2
3	22	32	-.23	.43	1.43	1.9	1.98	2.33	.07	.44	59.4	74.7	S3
8	22	32	-.23	.43	.89	-.5	.73	-.75	.55	.44	71.9	74.7	S8
5	23	32	-.42	.44	.74	-1.2	.59	-1.13	.64	.43	81.3	76.6	S5
6	24	32	-.62	.45	.80	-.8	.70	-.65	.57	.42	90.6	78.3	S6
9	24	32	-.62	.45	1.04	.2	.92	-.04	.40	.42	78.1	78.3	S9
1	26	32	-1.07	.50	1.00	.1	1.05	.28	.38	.39	81.3	83.3	S1
4	28	32	-1.64	.58	.72	-.7	.47	-.61	.55	.35	93.8	87.9	S4
MEAN	20.4	32.0	.00	.47	.99	.0	.96	.0			77.8	78.6	
P.SD	7.0	.0	1.31	.06	.21	.9	.40	.9			10.7	5.4	

Gambar 1. Hasil Analisis Validitas Butir Soal pada *Rasch* Model

Tabel 1 menunjukkan bahwa menurut teori tes klasik ditemukan lima butir soal yang dinyatakan pada kategori valid dan lima butir soal termasuk dalam kategori tidak valid. Lalu dalam analisis validitas soal oleh *Rasch* model ditemukan enam butir soal yang dinyatakan sebagai valid dan empat butir soal dinyatakan tidak valid karena tidak sesuai sedikitnya dua kriteria. Terdapat

butir soal yang tidak valid dikarenakan tidak memenuhi tiga kriteria, yaitu Soal#3. Untuk tiga soal lainnya dapat diperbaiki karena masih memenuhi satu kriteria. Ada perbedaan perolehan analisis dari dua pendekatan. Pengujian dengan *Rasch* model didapatkan soal dalam kategori valid lebih banyak dibandingkan tes teori klasik. Pendekatan *Rasch* model dianggap lebih benar karena analisis butir soal diperhatikan melalui tiga kriteria (*Outfit MNSQ*, *Outfit ZSTD*, serta *Pt Measure Corr*) maka butir soal tersebut dianggap valid. Jika soal tidak valid, harus diperhatikan konstruksi dari butir soal terkait adanya *stimulus, stem*, dan *option* (ada distraktor dan kunci jawaban).

Reliabilitas mengacu pada sejauh mana seperangkat pengukuran menghasilkan hasil yang konsisten. Reliabilitas berkaitan dengan konsistensi instrumen penelitian di berbagai pengaturan. Ini mencakup konsistensi hasil tes ketika dilaksanakan oleh individu yang berbeda (**inter-evaluator**), konsistensinya hasil tes saat dilaksanakan oleh orang yang sama pada waktu yang lain (pengulangan tes), konsistensinya hasil tes apabila dilaksanakan oleh individu yang berbeda secara simultan dengan memakai tes yang lain (bentuk paralel), serta konsisten hasilnya yang diperoleh dengan memakai kumpulan pertanyaan yang berbeda (konsistensi internal) (Jumini, 2023). Dengan begitu, reliabilitas adalah kualitas yang berhubungan dengan tingkat konsistensi yang ditunjukkan oleh kumpulan perangkat pengukuran. Keandalan elemen subjek atau reliabel suatu butir soal dalam pendekatan teori tes klasik dapat dinilai dengan nilai *Alpha Cronbach (KR-20)* (Sugiono, 2021). Butir soal dinyatakan dapat diandalkan atau reliabel apabila sesuai kriteria untuk koefisien korelasi keandalan instrumen yang dijelaskan dalam Tabel 2 (Erfan et al, 2020), seperti pada tabel di bawah ini. Nilai reliabilitas dari butir dan person melalui pendekatan *Rasch* model ditentukan berdasarkan kriteria yang disajikan dalam **Tabel 3** (Erfan et al., 2020).

Tabel 2. Kriteria Koefisien Korelasi Reliabilitas Instrumen pada Teori Tes Klasik

No	Koefisien Korelasi	Reliabilitas
1	r ≤ 0,20	Sangat rendah
2	0,20 ≤ r ≤ 0,40	Rendah
3	0,40 ≤ r ≤ 0,70	Sedang
4	0,70 ≤ r ≤ 0,90	Tinggi
5	0,90 ≤ r ≤ 1,00	Sangat tinggi

Tabel 3. Kriteria Reliabilitas Instrumen pada *Rasch* Model

No	Nilai Reliability (Person/Butir)	Interpretasi
1	>0,94	Istimewa
2	0,91 – 0,94	Bagus sekali
3	0,81 – 0,90	Bagus
4	0,67 – 0,80	Cukup
5	< 0,67	Lemah

Hasil analisis reliabilitas butir soal dengan tes teori klasik dan *Rasch* model ditunjukkan pada **Tabel 4**. Selain itu, untuk hasil butir soal dan person pada *Rasch* model disajikan dalam **Gambar 2**. Berdasarkan analisis teori klasik, reliabilitas instrumen diperoleh melalui perhitungan koefisien Alpha Cronbach, yang mencerminkan konsistensi internal dari seluruh butir soal. Sementara itu, pada *Rasch* model, reliabilitas ditunjukkan melalui nilai reliabilitas item dan person secara terpisah, yang menggambarkan seberapa tepat butir soal mampu membedakan kemampuan antar individu dan seberapa konsisten respon individu terhadap butir soal. **Gambar 2** juga menampilkan sebaran antara tingkat kesulitan soal dan kemampuan siswa dalam satu peta (person-item map), yang dapat digunakan untuk mengevaluasi keseimbangan instrumen, apakah sudah sesuai dengan karakteristik peserta tes. Dengan demikian, kedua pendekatan ini saling melengkapi dalam memberikan gambaran menyeluruh terhadap kualitas reliabilitas instrumen yang digunakan dalam penelitian.

Tabel 4. Hasil Analisis Reliabilitas Soal dengan Cara Klasik dan *Rasch* Model

Teori Tes Klasik		***Rasch* Model**	
Kriteria Kategori Alpha Cronbach = 0,458	Sedang	Kriteria Person Reliability = 0,34	Lemah
		Kriteria Item Reliability = 0,86	Bagus

File Edit Format View Help

SUMMARY OF 32 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	6.4	10.0	.76	.85	.99	.02	.96	.05
SEM	.3	.0	.21	.03	.08	.17	.12	.15
P.SD	1.8	.0	1.16	.15	.45	.95	.64	.83
S.SD	1.8	.0	1.18	.15	.46	.96	.65	.84
MAX.	9.0	10.0	2.93	1.24	2.20	2.04	2.54	2.07
MIN.	2.0	10.0	-1.77	.71	.39	-1.40	.13	-.95

REAL RMSE	.94	TRUE SD	.68	SEPARATION	.72	Person RELIABILITY .34
MODEL RMSE	.86	TRUE SD	.78	SEPARATION	.90	Person RELIABILITY .45

S.E. OF Person MEAN = .21

Person RAW SCORE-TO-MEASURE CORRELATION = .99

CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .48 SEM = 1.30

SUMMARY OF 10 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	20.4	32.0	.00	.47	.99	.04	.96	.05
SEM	2.3	.0	.44	.02	.07	.30	.13	.31
P.SD	7.0	.0	1.31	.06	.21	.90	.40	.93
S.SD	7.3	.0	1.38	.06	.22	.95	.42	.98
MAX.	28.0	32.0	3.13	.58	1.43	1.97	1.98	2.33
MIN.	4.0	32.0	-1.64	.41	.72	-1.27	.47	-1.13

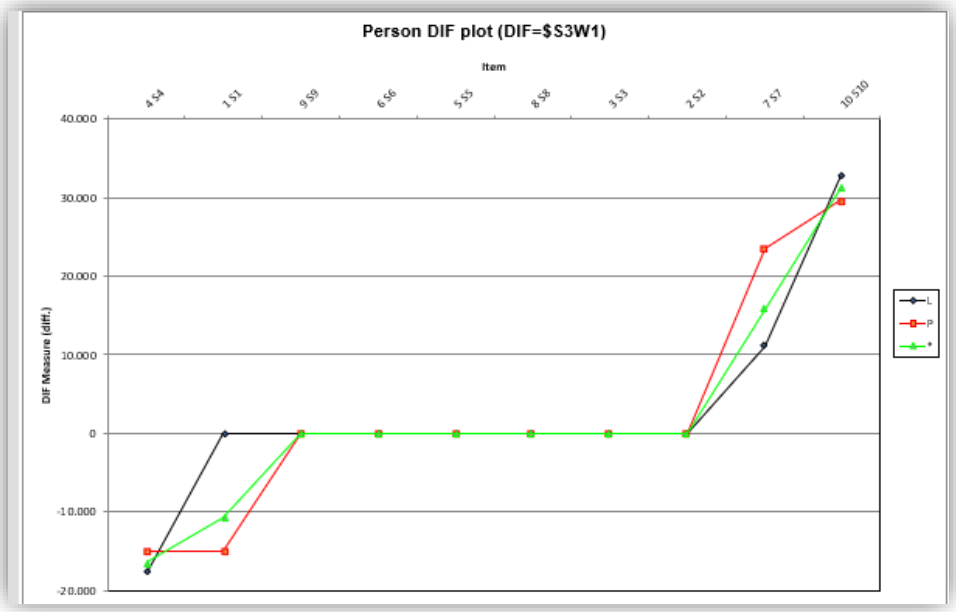
REAL RMSE	.49	TRUE SD	1.22	SEPARATION	2.50	Item RELIABILITY .86
MODEL RMSE	.47	TRUE SD	1.22	SEPARATION	2.59	Item RELIABILITY .87

S.E. OF Item MEAN = .44

Item RAW SCORE-TO-MEASURE CORRELATION = -1.00

Gambar 2. Hasil Analisis Reliabilitas Pearson dan Item Soal dan pada *Rasch* Model

Menurut teori tes klasik (**Tabel 4**), reliabilitas pada keseluruhan bernilai konsisten sedang (Alpha cronbach= 0,458). Sementara itu, skor reliabilitas person 0,34 pada *Rasch* model bernilai lemah. Hal ini dapat disebabkan peserta didik dalam menjawab pertanyaan tidak konsisten. Dengan kata lain, reliabilitas peserta didik yang lemah mengacu pada hasil belajar mereka tidak konsisten dari waktu ke waktu atau dari satu situasi ke situasi lain. Selain itu, untuk kualitas dari person jika dilihat dari DIF (**Gambar 3**), terdapat perbedaan yang signifikan dari soal nomor 1 yang cenderung dapat dijawab benar oleh peserta didik dengan jenis kelamin laki-laki daripada perempuan, kemudian soal nomor 7 yang cenderung dapat dijawab benar oleh peserta didik dengan jenis kelamin perempuan daripada laki-laki. Namun, hasil ini tidak signifikan karena kurang dari 0,05 atau 5% (probabilitas). Nilai reliabilitas item soal sebesar 0,86 menunjukkan kategori bagus. Hasil ini menunjukkan kualitas butir-butir soal yang reliabilitas untuk membantu peserta didik belajar sehingga dapat meningkatkan kualitas, mendorong pembelajaran, dan mendukung penilaian yang akurat. Hal ini sesuai dengan Tarigan (2022), bahwa perangkat pengukuran (instrumen) yang reliabel akan menghasilkan data yang lebih dapat diandalkan dan akurat, memungkinkan pengeluaran keputusan yang unggul. Item soal yang reliabel artinya jawaban peserta didik yang memiliki kemampuan yang sama akan selalu sama, meskipun soal tersebut diujikan pada waktu yang berbeda atau dengan kondisi yang berbeda.



Gambar 3. Hasil Analisis DIF Person pada *Rasch* Model

Tingkatan kesukaran butir soal adalah ukuran yang memperlihatkan kemungkinan berapa banyaknya peserta didik yang dapat menjawab pertanyaan dengan benar (Fauziana & Dessy, 2021). Tingkat kesukaran item menentukan apakah itu jatuh pada kategori sukar, sedang, atau mudah. Soal yang baik ditandai oleh tingkatan kesukaran yang tidak terlalu susah atau terlalu mudah. Semakin sukar soal, semakin tinggi tingkat kesulitannya. sedangkan soal yang terlalu sukar dapat membuat mereka frustasi dan kehilangan minat. Hal ini berarti soal itu membutuhkan lebih banyak pengetahuan, keterampilan, dan kemampuan. Soal yang terlalu mudah tidak akan menantang peserta didik untuk belajar dan berkembang Tingkat kesulitan soal yang ideal adalah yang cukup menantang untuk mendorong peserta didik belajar dan berkembang, tetapi tidak terlalu sulit sehingga membuat mereka frustasi. Tingkat kesukaran suatu butir soal dalam teori tes klasik dihitung berdasarkan kriteria yang dijelaskan dalam **Tabel 5** (Erfan et al., 2020).

Tabel 5. Kriteria Tingkat Kesukaran Butir Soal pada Tes Klasik

No	Nilai Indeks Kesukaran	Interpretasi
1	IK=0,00	Sangat sukar
2	0,00 < IK ≤ 0,30	Sukar
3	0,30 < IK ≤ 0,70	Sedang
4	0,70 < IK ≤ 1,00	Mudah
5	IK=1,00	Terlalu mudah

Menurut *Rasch* model, ada empat kelompok atau kategori untuk tingkat kesukaran item berdasarkan Measure Logit dan Simpang Baku (SD) dari item logit sebagai berikut **Tabel 6** (Erfan et al., 2020). Hasil analisis tingkat kesukaran dengan menggunakan tes teori klasik dan *Rasch* model disajikan dalam **Tabel 7**.

Tabel 6. Kriteria Tingkat Kesukaran Butir Soal dengan Pemodelan *Rasch*

No	Nilai Measure (logit)	Interpretasi
1	Measure logit > SD logit	Sangat sukar
2	0 ≤ Measure logit ≤ SD logit	Sukar
3	0 ≤ Measure logit ≤ 0	Sedang
4	Measure logit < - SD logit	Mudah

Tabel 7. Hasil Analisis Tingkat Kesukaran Soal dengan Tes Teosi Klasik dan *Rasch* Model

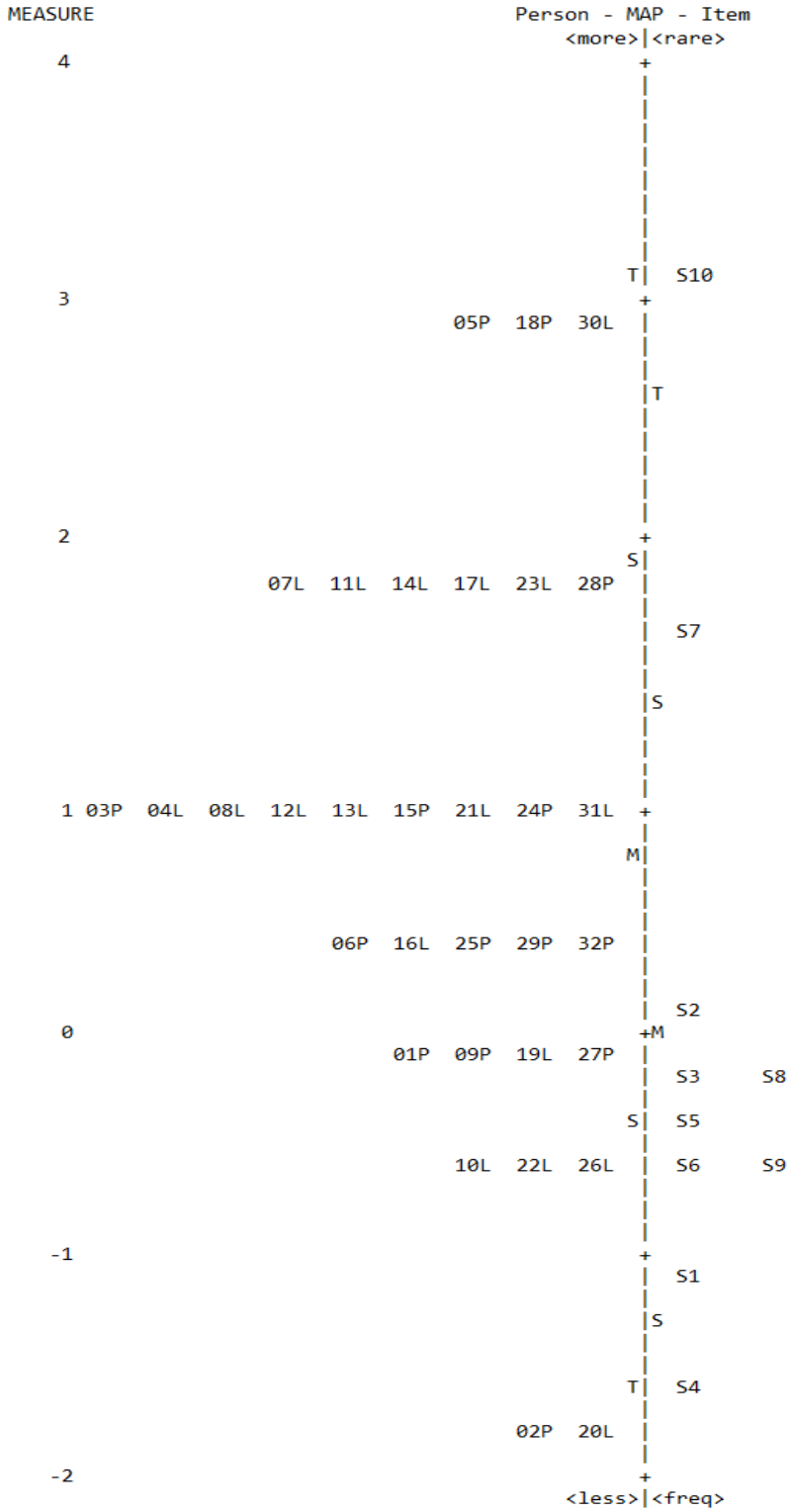
No	Kriteria	No Soal	
		Teori Tes Klasik	***Rasch* Model**
1.	Sangat sukar	-	-
2.	Sukar	10	10
3.	Sedang	2,3,7,8	1,2,3,5,6,7,8,9
4.	Mudah	1,4,5,6,9	4

Berdasarkan **Tabel 6** dan **Tabel 7**, tingkat kesukaran butir soal dianalisis menggunakan dua pendekatan, yaitu teori tes klasik dan pemodelan *Rasch*. Kriteria pengelompokan tingkat kesukaran menurut *Rasch* didasarkan pada nilai measure (logit), yang dibagi menjadi empat kategori, yaitu sangat sukar jika nilai logit lebih besar dari simpangan baku (SD), sukar jika berada antara 0 hingga SD, sedang jika berada antara 0 hingga nilai negatif, dan mudah jika lebih kecil dari negatif SD. Hasil analisis dalam **Tabel 7** menunjukkan bahwa tidak terdapat butir soal yang termasuk kategori sangat sukar, baik menurut teori tes klasik maupun model *Rasch*.

Dalam kategori sukar, baik teori tes klasik maupun *Rasch* model sama-sama mengidentifikasi soal nomor 10 sebagai soal yang tergolong sukar. Untuk kategori sedang, teori tes klasik mencakup soal nomor 2, 3, 7, dan 8, sedangkan *Rasch* model mencakup soal yang lebih luas, yaitu nomor 1, 2, 3, 5, 6, 7, 8, dan 9. Hal ini menunjukkan bahwa model *Rasch* memberikan distribusi yang lebih merata dalam mengelompokkan tingkat kesukaran butir soal. Sementara itu, untuk kategori mudah, teori tes klasik mengelompokkan soal nomor 1, 4, 5, 6, dan 9, sedangkan *Rasch* model hanya mengidentifikasi soal nomor 4 sebagai soal mudah. Perbedaan ini menggambarkan bahwa pendekatan *Rasch* memiliki sensitivitas yang lebih tinggi dalam menilai kesukaran butir soal karena mempertimbangkan interaksi antara kemampuan peserta dan kesukaran soal secara simultan.

Dengan demikian, hasil analisis dari kedua pendekatan ini menunjukkan adanya perbedaan dalam pengelompokan tingkat kesukaran butir soal. Penggunaan pemodelan *Rasch* memberikan informasi yang lebih mendalam dan objektif, sehingga sangat membantu dalam menyusun instrumen yang tepat sasaran dan seimbang sesuai dengan kemampuan peserta didik.

Hasil analisis dengan *Rasch* model membagi item menjadi tiga kelompok, yaitu 1 soal mudah, 8 soal sedang, dan 1 soal pertanyaan sukar. *Rasch* model memberikan informasi tentang tingkat kesulitan dari item soal pengukuran, seperti yang dapat diamati dalam **Gambar 4**.



Gambar 4. Hasil Analisis Item Map Tingkat Kesukaran pada *Rasch* Model

Wright map pada **Gambar 4** menunjukkan bahwa soal nomor 10 memiliki nilai logit tertinggi dengan kategori sukar, dan tidak ada satupun peserta didik yang mampu mengerjakan dengan benar. Urutan dibawahnya soal nomor 1,2,3,5,6,7,8,9 kategori sedang, dimana sebagian besar peserta didik dapat mengerjakan soal dalam kategori ini dengan benar. Soal urutan paling bawah nomor 4 kategori mudah, hampir semua peserta didik mampu menjawab benar namun terdapat 2 peserta

didik yang masih menjawab salah. Hal ini sesuai dengan buku oleh Rahmi (2022), menjelaskan bahwa tingkat kesukaran subjek dalam teori klasik tergantung pada komposisi peserta didik.

Daya pembeda adalah kemampuan dalam menyaring data guna membedakan di antara peserta didik tingkat tinggi dan rendah dilihat dari kemampuan dalam menjawab butir soal dengan benar. Moeriyandani (2021) mendefinisikan kekuatan diferensial sebagai metrik yang mengevaluasi kemampuan seseorang untuk membedakannya diantara seorang siswa yang telah mencapai penguasaan subjek dan seorang siswa yang belum mencapai. Hasil analisis daya pembeda pada butir soal melalui teori tes klasik, ditampilkan dalam **Tabel 8** (Erfan et al., 2020).

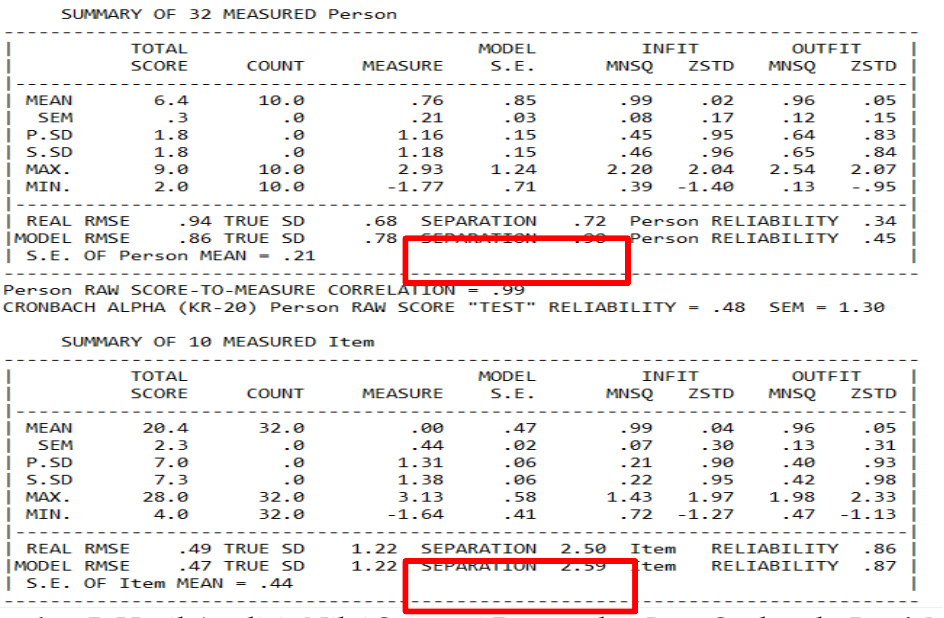
Tabel 8. Hasil Analisis Daya Beda Butir Soal dengan Teori Tes Klasik

Daya Pembeda	Interpretasi	Jumlah Butir Soal
DP ≥ 0,70	Sangat Tinggi	1
0,40 ≤ DP < 0,70	Tinggi	4
0,20 ≤ DP < 0,40	Cukup	5
DP < 0,20	Rendah	0

Hasil analisis dengan teori tes klasik menunjukkan bahwa data beda 5 soal bernilai cukup, 4 soal bernilai tinggi, dan 1 soal bernilai sangat tinggi. Uji *Rasch* model digunakan guna membedakannya diantara peserta didik yang dapat menjawab pertanyaan dan mereka yang kurang mampu melakukannya. Nilai Model Standar Error (SE) memberikan cara untuk mengamati tingkat daya pembeda (Meisya, 2023). Nilai model SE di bawah 0,5 menunjukkan kekuatan daya pembeda yang bagus atau kuat dari item. Jika nilai jatuh antara 0,5 dan 1, perbedaan dianggap cukup untuk dapat dibedakan. Nilai di atas 1 menunjukkan kekuatan perbedaan yang buruk atau ketidakmampuan untuk membedakan (Purniasari et al, 2021). Cara menentukan kelompok peserta didik dengan menggunakan indeks separasi person dan item soal. Kualitas instrumen meningkat seiring dengan peningkatan nilai separasi item, karena memungkinkan untuk mengidentifikasi kedua kelompok person dan item soal yang dianalisis. Persamaan lain yang memerlukan pengelompokan yang teliti adalah persamaan *strata (H)*.

$$H = \frac{(4\ x\ Separation)+1}{3} \qquad (1)$$

Hasil analisis daya pembeda person dan item ditunjukkan pada **Gambar 5**. Hasil menunjukkan bahwa skor separasi item soal sebanyak 2,50 dengan skor H = 3,67 pembulatan jadi 4, hingga ada empat kelompok butir soal yang bisa dikaji. Sementara itu nilai separasi person sebesar 0,72 dengan skor H = 1,29 pembulatan menjadi 1, memperlihatkan bahwa terdapat satu kelompok kemampuan peserta didik. Perolehan analisis daya pembeda butir soal secara klasik dan modern didapat hasil yang berbeda. Nilai separasi pada tes teori klasik dibagi tiga (sangat tinggi, tinggi, dan cukup) sedangkan *Rasch* model dibagi empat (sangat tinggi, tinggi, cukup, dan rendah) (Novia, 2019). Dengan begitu, pendidik dapat memastikan bahwa soal yang digunakan dapat membedakan dengan baik antara peserta didik yang pandai dan peserta didik yang kurang pandai, sehingga hasil tes atau ujian yang diperoleh lebih akurat dan terpercaya.



SUMMARY OF 32 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	6.4	10.0	.76	.85	.99	.02	.96	.05
SEM	.3	.0	.21	.03	.08	.17	.12	.15
P.SD	1.8	.0	1.16	.15	.45	.95	.64	.83
S.SD	1.8	.0	1.18	.15	.46	.96	.65	.84
MAX.	9.0	10.0	2.93	1.24	2.20	2.04	2.54	2.07
MIN.	2.0	10.0	-1.77	.71	.39	-1.40	.13	-.95

REAL RMSE	.94	TRUE SD	.68	SEPARATION	.72	Person RELIABILITY .34
MODEL RMSE	.86	TRUE SD	.78	SEPARATION	.90	Person RELIABILITY .45
S.E. OF Person MEAN = .21						

Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .48 SEM = 1.30

SUMMARY OF 10 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	20.4	32.0	.00	.47	.99	.04	.96	.05
SEM	2.3	.0	.44	.02	.07	.30	.13	.31
P.SD	7.0	.0	1.31	.06	.21	.90	.40	.93
S.SD	7.3	.0	1.38	.06	.22	.95	.42	.98
MAX.	28.0	32.0	3.13	.58	1.43	1.97	1.98	2.33
MIN.	4.0	32.0	-1.64	.41	.72	-1.27	.47	-1.13

REAL RMSE	.49	TRUE SD	1.22	SEPARATION	2.50	Item RELIABILITY .86
MODEL RMSE	.47	TRUE SD	1.22	SEPARATION	2.59	Item RELIABILITY .87
S.E. OF Item MEAN = .44						

Gambar 5. Hasil Analisis Nilai Separasi Person dan Item Soal pada *Rasch* Model

Efisiensi dari distraktor ditentukan oleh sejauh mana pilihan jawaban yang salah (distraktor) dalam pertanyaan pilihan ganda dapat mengecoh peserta didik yang tidak menyadari jawaban yang benar (Simamora, 2021). Distraktor yang efektif akan menarik peserta didik yang kurang dalam memahami materi, sehingga memaksa mereka untuk memilihnya. Jika jumlah peserta didik yang memilih suatu distraktor meningkat, efektivitasnya distraktor tersebut juga meningkat. Distraktor yang efektif dapat meningkatkan kualitas soal dan membantu untuk membedakan mana peserta didik yang memahami materi dengan yang tidak (Rustam, 2023). Hasil uji efektivitas distraktor tes teori klasik disajikan dalam **Tabel 9**. Hasil menunjukkan distraktor yang paling baik terdapat pada nomor 4 dengan pilihan jawaban A sebesar 88%, sedangkan beberapa distraktor tidak bekerja karena bernilai 0%. Hasil uji efektivitas distraktor dengan *Rasch* model ditampilkan pada **Gambar 6**. Hasil menunjukkan ada beberapa distraktor yang tidak dipilih siswa sehingga bernilai buruk atau bermasalah. Lalu, dalam distraktor yang terpilih harus selalu bernilai tinggi atau naik dalam kemampuannya.

Tabel 9. Efektivitas Distraktor dari Pilihan Jawaban Soal secara Teori Tes Klasik

No	Efektivitas Distraktor			
	A	B	C	D
1.	0%	-	19%	0%
2.	-	25%	13%	0%
3.	22%	0%	9%	-
4.	-	0%	0%	13%
5.	3%	25%	-	0%
6.	16%	0%	-	9%
7.	-	22%	44%	0%
8.	-	9%	9%	13%
9.	13%	-	13%	0%
10.	28%	-	50%	9%

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
10	D	0	3	9	-.13	.42	.30	.2	.3	-.25	S10
	A	0	9	28	.30	1.30	.46	.6	.7	-.25	
	C	0	16	50	.99	1.02	.26	1.3	1.4	.20	
	B	1	4	13	1.52	.95	.55	1.2	1.1	.25	
7	C	0	14	44	.30	.94	.26	.8	.8	-.35	S7
	B	0	7	22	.57	1.15	.47	1.1	1.1	-.09	
	A	1	11	34	1.46	1.09	.35	1.1	1.1	.44	
2	C	0	4	13	.15	.61	.35	.8	.7	-.20	S2
	B	0	8	25	.32	1.06	.40	1.2	1.1	-.22	
	A	1	20	63	1.05	1.19	.27	1.3	1.4	.33	
3	A	0	7	22	.46	.77	.32	1.2	1.2	-.14	S3
	C	0	3	9	1.06	1.34	.95	1.8	4.0	.08	
	D	1	22	69	.81*	1.22	.27	1.5	1.8	.07	
8	D	0	4	13	-.42	.81	.47	.7	.5	-.38	S8
	B	0	3	9	-.13	1.19	.84	1.1	.8	-.25	
	C	0	3	9	.07	.68	.48	.9	.8	-.19	
	A	1	22	69	1.18	1.00	.22	.9	.9	.55	
5	B	0	8	25	-.61	.74	.28	.6	.4	-.68	S5
	A	0	1	3	.99	.00		2.1	1.6	.04	
	C	1	23	72	1.22	.91	.19	.7	.7	.64	
6	A	0	5	16	-.44	1.24	.62	.9	.8	-.44	S6
	D	0	3	9	-.30	.48	.34	.7	.5	-.29	
	C	1	24	75	1.14	.94	.20	.7	.7	.57	
9	C	0	4	13	-.39	1.38	.80	1.1	.9	-.37	S9
	A	0	4	13	.28	.59	.34	1.2	1.0	-.15	
	B	1	24	75	1.03	1.05	.22	1.0	.9	.40	
1	C	0	6	19	-.17	1.09	.49	1.0	1.1	-.38	S1
	B	1	26	81	.97	1.07	.21	1.1	1.0	.38	
4	D	0	4	13	-.95	.90	.52	.6	.4	-.55	S4
	A	1	28	88	1.00	.98	.19	.6	.8	.55	

* Average ability does not ascend with category score

Gambar 6. Hasil Analisis Efektivitas Distraktor Pilihan Jawaban Soal pada *Rasch* Model

KESIMPULAN

Pada perolehan analisis instrumen penilaian kualitas soal pemahaman materi keanekaragaman hayati pada 32 peserta didik dari salah satu kelas VII di SMP Negeri daerah Sukoharjo melalui pendekatan teori tes klasik diklasifikasikan menjadi soal kualitas cukup ataupun sedang dari segi validitas butir soal (5 butir soal valid dari total 10 butir soal) lalu dengan analisis *Rasch* model diperoleh kualitas soal yang baik (6 butir soal valid dari total 10 butir soal). Analisis reliabilitas butir soal lewat pendekatan teori tes klasik didapati skor Alpha Cronbach untuk skor reliabilitas sejumlah 0,458 pada kategori sedang dan melalui *Rasch* model diperoleh person bernilai lemah (0,34) dan butir soal bernilai bagus (0,86). Dalam tingkat kesukaran, berlandaskan teori tes klasik terbagi pada tiga kelompok, yaitu mudah 5 soal, sedang 4 soal, dan sukar 1 soal. Sedangkan pada *Rasch* model, butir soal terdistribusi dalam tiga kelompok juga, yakni mudah 1 soal, sedang 8 soal, dan sukar 1 soal. Setelah dianalisis dengan pendekatan teori tes klasik, sebagian besar butir soal memperlihatkan dalam daya pembeda di 5 soal bernilai cukup, 4 soal bernilai tinggi, dan 1 soal bernilai sangat tinggi. Lalu, pada *Rasch* model terdapat empat kelompok butir soal yang dapat diidentifikasi, sedangkan untuk person hanya terdapat satu kelompok. Untuk efektivitas distraktor pada tes teori klasik diperoleh hasil bahwa distraktor yang paling baik terdapat pada nomor 4 dengan pilihan jawaban A sebesar 88%, sedangkan beberapa distraktor tidak bekerja karena bernilai 0%. Dalam hasil *Rasch* model, diperoleh beberapa distraktor yang tidak ada yang memilih sehingga beberapa pilihan jawaban bernilai buruk atau bermasalah. Lalu, dalam distraktor yang terpilih harus selalu bernilai tinggi atau naik dalam ability nya. Dalam analisis pendekatan melalui dua cara ini, pendekatan *Rasch* model dianggap lebih baik dikarenakan lebih objektif, mudah dalam menafsirkan hasil, fleksibel, dan kuat secara statis. Dalam analisis instrumen penilaian untuk menganalisis butir soal dapat menggunakan lebih banyak pendekatan agar ketepatan dalam analisis butir soal dapat lebih berkualitas dan diketahui dimana letak kesalahan, kekurangan, dan ketidaksempurnaan dari butir soal yang diujikan kepada peserta didik.

DAFTAR PUSTAKA

Antara, A. A. P., Yasna, I. M., Dewi, N. W. D. P., & Maduriana, I. M. (2019). Karakteristik Tes Prestasi Belajar Model Campuran Dikotomus dan Politomus Generalized Partial Credit Model (GPCM). *Suluh Pendidikan: Jurnal Ilmu-Ilmu Pendidikan*, *17*(1), 83-94.

Arifin, Z. (2022). Manajemen Peserta Didik sebagai Upaya Pencapaian Tujuan Pendidikan. *Dirasat: Jurnal Manajemen dan Pendidikan Islam*, *8*(1), 71-89. https://doi.org/10.26594/dirasat.v8i1.3025

Azwar, S. (2019). *Penyusunan Skala Psikologi*. Yogyakarta: Pustaka Pelajar

Bichi, A. A., Embong, R., Talib, R., Salleh, S., & Bin Ibrahim, A. (2019). Comparative Analysis of Classical Test Theory and Item Response Theory using Chemistry Test Data. *International Journal of Engineering and Advanced Technology*, *8*(5), 1260-1266. https://doi.org/10.35940/ijeat.E1179.0585C19

Bungin, B. (2021). *Metode Penelitian Sosial*. Universitas Airlangga Press.

Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). 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.

Fernando, D. A., Hartatiana, H., & Ismail, F. (2023). Pentingnya Validitas dan Reliabilitas Instrumen Evaluasi pada Pembelajaran Pendidikan Agama Islam. *Raudhah Proud to Be Professionals: Jurnal Tarbiyah Islamiyah*, *8*(3), 1110-1121. https://doi.org/10.48094/raudhah.v8i3.580

Firmansyah, D. (2022). Teknik Pengambilan Sampel Umum dalam Metodologi Penelitian: Literature Review. *Jurnal Ilmiah Pendidikan Holistik (JIPH), 1*(2), 85-114.

Guangul, F. M., Suhail, A. H., Khalit, M. I., & Khidhir, B. A. (2020). Challenges of Remote Assessment in Higher Education in The Context Of COVID-19: A Case Study of Middle East College. *Educational Assessment, Evaluation And Accountability*, *32*, 519-535. https://doi.org/10.1007/s11092-020-09340-w

Habibi, M., Lestari, F. A., & Afif, Y. U. (2021). Implementasi Penilaian Autentik Kurikulum 2013 Pada Mata Pelajaran PAI dan Budi Pekerti Di SDN 1 Bangunrejo Ponorogo. *QALAMUNA: Jurnal Pendidikan, Sosial, Dan Agama*, *13*(2), 833-852. https://doi.org/10.37680/qalamuna.v13i2.1114

Hardianti, H. (2021). Karakteristik Tes Kemampuan Berpikir Kritis Siswa SMA pada Materi Momentum dan Impuls: Perbandingan Classical Theory Test (CTT) dan Model *Rasch*. *WaPFi (Wahana Pendidikan Fisika), 6*(2), 167–173. https://doi.org/10.17509/wapfi.v8i1.30958

Ida, F. F., & Musyarofah, A. (2021). Validitas dan Reliabilitas dalam Analisis Butir Soal. *Al-Mu'arrib: Jurnal Pendidikan Bahasa Arab*, *1*(1), 34-44.

Izza, A. Z., Falah, M., & Susilawati, S. (2020). Studi literatur: Problematika Evaluasi Pembelajaran dalam Mencapai Tujuan Pendidikan di Era Merdeka Belajar. *Prosiding Konferensi Ilmiah Pendidikan*, *1*, 10-15.

Jumini, S., Madnasri, S., Cahyono, E., & Parmin, P. (2023, June). Analisis Kualitas Butir Soal Pengukuran Literasi Sains Melalui Teori Tes Klasik Dan *Rasch* Model. In *Prosiding Seminar Nasional Pascasarjana* (Vol. 6, No. 1, pp. 758-765).

Laliyo, L. A. R. (2021). *Mendiagnosis Sifat Perubahan Konseptual Siswa: Penerapan Teknik Analisis Stacking dan Racking Rasch Model*. Deepublish.

Magdalena, I., Hidayati, N., Dewi, R. H., Septiara, S. W., & Maulida, Z. (2023). Pentingnya Evaluasi dalam Proses Pembelajaran dan Akibat Memanipulasinya. *Masaliq*, *3*(5), 810-823. https://doi.org/10.36088/bintang.v2i2.986

Meisya, R., Jannah, R., & Ramadhan, S. (2023). Analisis Kualitas Butir Soal Tematik Madrasah Ibtidaiyah Menggunakan Model *Rasch*. *Al-Madrasah: Jurnal Pendidikan Madrasah Ibtidaiyah,* 7(4), 1764. https://doi.org/10.35931/am.v7i4.2712.

Moeriyandani, R., & Yulianto, B. (2021). Analisis Butir Tes Dokkai Siswa Kelas XI Peminatan Bahasa Jepang di SMA Negeri 1 Sumenep. *Jurnal Education and Development*, *9*(3), 461-466. https://doi.org/10.17509/wapfi.v8i1.30958

Nasution, I., Harahap, H. T., Nurfadillah, L., & Purba, S. L. B. (2022). Evaluasi Program Ekstrakurikuler pada Sekolah MIS Nur Al Amin Medan. *Jurnal Pendidikan dan Konseling (JPDK)*, *4*(3), 1638-1646. https://doi.org/10.31004/jpdk.v4i3.4931

Novia, R., Ramalis, T. R., & Efendi, R. (2019). Pengembangan dan Karakterisasi Tes Keterampilan Berpikir Kritis Materi Tekanan berdasarkan Teori Respon Butir. *WaPFi (Wahana Pendidikan Fisika)*, 4(2), 155–162. https://doi.org/10.17509/wapfi.v4i2.20181.

Polat, M., Turhan, N. S., & Toraman, Ç. (2022). Comparison of Classical Test Theory vs. Multi-Facet *Rasch* Theory in Writing Assessment. *Pegem Journal of Education and Instruction*, *12*(2), 213-225. https://doi.org/10.47750/pegegog.12.02.21

Pramana, K. A. B., & Putra, D. B. K. N. S. (2019). *Merancang Penilaian Autentik*. Cv. Media Educations.

Pratiwi, L. E. P., & Rufi'i, H. (2023). Analisis Soal Sumatif IPA Mengenai Gaya pada Peserta Didik Kelas IV di SD YP Nasional. *Didaktika: Jurnal Kependidikan*, *12*(4), 599-610. https://doi.org/10.58230/27454312.287

Pudjiati, I., & Madani, F. (2023). Asesmen Autentik Analisis Butir Soal Dengan *Rasch* Model Di Sekolah Dasar: Literature Review. *Mutiara: Jurnal Penelitian dan Karya Ilmiah*, *1*(4), 01-12. https://doi.org/10.59059/mutiara.v1i4.325

Purniasari, L., Masykuri, M., & Ariani, S. R. D. (2021). Analisis Butir Soal Ujian Sekolah Mata Pelajaran Kimia Sma N 1 Kutowinangun Tahun Pelajaran 2019/2020 Menggunakan Model Iteman dan *Rasch*. *Jurnal Pendidikan Kimia, 10*(2), 205–214. https://doi.org/10.47750/pegegog.12.02.21

Purwaningsih, E., & Suryadi, A. (2022). *Penelitian Kuantitatif Pendidikan Fisika (Topik, Instrumen, dan Statistik Dasar)*. Bayfa Cendekia Indonesia.

Puspita, V., & Dewi, I. P. (2021). Efektifitas E-LKPD berbasis Pendekatan Investigasi terhadap Kemampuan Berpikir Kritis Siswa Sekolah Dasar. *Jurnal Cendekia: Jurnal Pendidikan Matematika*, *5*(1), 86-96. https://doi.org/10.31004/cendekia.v5i1.456

Rahmi, M. A., Kustati, M., & Hadeli, M. A. (2022). *Evaluasi Pendidikan Perspektif Islam*. Deepublish.

Rustam, A., Iriyadi, D., Ekadayanti, W., & Info, A. (2023). Basic Algebra Ability Analysis of Junior High School students with the *Rasch* Model Approach. *JME (Journal of Mathematics Education)*, 8(1), 47–55. https://doi.org/10.31327/jme.v8i1.1908.

Simamora, H., Hartono, H., & Effendi, E. (2021). Analisis Kualitas Butir Soal Buatan Guru Kimia pada Tes Ujian Tengah Semester Ganjil Kelas XII MIPA. *Hydrogen: Jurnal Kependidikan Kimia*, *9*(1), 8-18. https://doi.org/10.33394/hjkk.v9i1.3701

Suardipa, I. P., & Primayana, K. H. (2023). Peran Desain Evaluasi Pembelajaran untuk Meningkatkan Kualitas Pembelajaran. *Widyacarya: Jurnal Pendidikan, Agama dan Budaya*, *4*(2), 88-100. https://doi.org/10.55115/widyacarya.v4i2.796

Sugiyono. (2021). *Metode Penelitian Kualitatif*. Bandung : Alfabeta

Suratman, A., Afyaman, D., & Rakhmasari, R. (2019). Pembelajaran Berbasis TIK terhadap Hasil Belajar Matematika dan Motivasi Belajar Matematika Siswa. *Jurnal Analisa*, *5*(1), 41-50. https://doi.org/10.15575/ja.v5i1.4828

Sylvia, I., Anwar, S., & Khairani, K. (2019). Pengembangan Instrumen Penilaian Autentik Berbasis Pendekatan Authentic Inquiry Learning Pada Mata Pelajaran Sosiologi di Sekolah Menengah Atas. *Jurnal Socius: Journal of Sociology Research and Education*, *6*(2), 103-120. https://doi.org/10.24036/scs.v6i2.162

Tamrin, M. I. (2019). Peningkatan Sumber Daya Manusia dalam Lembaga Pendidikan Agama Non Formal di Era Global. *Menara Ilmu: Jurnal Penelitian dan Kajian Ilmiah*, *13*(2, 94-101). https://doi.org/10.58230/27454312.287

Tarigan, E. F., Nilmarito, S., Islamiyah, K., Darmana, A., & Suyanti, R. D. (2022). Analisis Instrumen Tes Menggunakan *Rasch* Model dan Software SPSS 22.0. *Jurnal Inovasi Pendidikan Kimia*, *16*(2), 92-96. https://doi.org/10.15294/jipk.v16i2.30530

Tarmizi, P., Setiono, P., Amaliyah, Y., & Agrian, A. (2020). Analisis butir soal pilihan ganda tema sehat itu penting kelas V SD Negeri 04 Kota Bengkulu. *ELSE (Elementary School Education Journal): Jurnal Pendidikan dan Pembelajaran Sekolah Dasar*, *4*(2), 124-132. https://doi.org/10.30651/else.v4i2.7090