

Asesmen Formatif: Analisis Psikometrik Instrumen Menggunakan Model Rasch Pada Materi Teorema Pythagoras

Irmawaty Natsir^{1*}, Anis Munfarikhatin¹, Khumaeroh Dwi Nur'Aini¹, Nurhayati²,
Nurmayaningsih³, Wiwik Andriani³

¹Pendidikan Matematika, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Musamus,
Merauke, Indonesia

²Pendidikan Komputer, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Musamus,
Merauke, Indonesia

³Pendidikan Teknologi Informasi, Fakultas Keguruan dan Ilmu Pendidikan, Universitas
Megarezky, Makassar, Indonesia

*Penulis Korespondensi: natsir_fkip@unmus.ac.id

Received: 11-05-2026; Revised: 11-06-2026; Published: 30-06-2026

Abstract: This study uses a quantitative approach that aims to analyze the psychometric characteristics of formative assessment instruments using the Rasch Model. The research sample consisted of 20 students. The instrument used was a multiple-choice test consisting of 20 questions on the Pythagorean Theorem. Data were analyzed using Winsteps to identify instrument reliability, item difficulty level, item fit, person measure, and student response patterns. The results showed that the instrument had a fairly good reliability with a Cronbach Alpha value of 0.70, person reliability of 0.67, and item reliability of 0.77. Analysis of item difficulty levels showed that there were 2 items in the very easy category, 7 items in the easy category, 8 items in the difficult category, and 3 items in the very difficult category. The results of the item fit analysis showed that most of the items were in accordance with the Rasch Model, although there is one item that needs to be revised because it does not meet the model suitability criteria. Analysis of student abilities showed that most students were in the medium to high ability category. In addition, the results of the Wright Map and Scalogram showed that the instrument had a fairly good ability to differentiate student abilities. Thus, the developed formative assessment instrument can be used to measure students' abilities in the Pythagorean Theorem material.

Keywords: formative assessment, psychometric characteristics, instrument quality, Rasch model

Abstrak: Penelitian ini menggunakan pendekatan kuantitatif yang bertujuan untuk menganalisis karakteristik psikometrik instrumen asesmen formatif menggunakan Model Rasch. Sampel penelitian terdiri atas 20 siswa. Instrumen yang digunakan berupa tes pilihan ganda sebanyak 20 butir soal pada materi Teorema Pythagoras. Data dianalisis menggunakan Winsteps untuk mengidentifikasi reliabilitas instrumen, tingkat kesulitan butir soal, item fit, person measure, serta pola respon siswa. Hasil penelitian menunjukkan bahwa instrumen memiliki reliabilitas yang cukup baik dengan nilai Cronbach Alpha sebesar 0,70, person reliability sebesar 0,67 dan item reliability sebesar 0,77. Analisis tingkat kesulitan butir soal menunjukkan terdapat 2 butir dengan kategori sangat mudah, 7 butir dengan kategori mudah, 8 butir dengan kategori sulit, dan 3 butir dengan kategori sangat sulit. Hasil analisis item fit menunjukkan bahwa sebagian besar butir soal sesuai dengan Model Rasch, meskipun terdapat satu butir yang perlu direvisi karena belum memenuhi kriteria kesesuaian model. Analisis kemampuan siswa menunjukkan bahwa sebagian besar siswa berada pada kategori kemampuan sedang hingga tinggi. Selain itu, hasil *Wright Map* dan *Scalogram* menunjukkan bahwa instrumen memiliki kemampuan yang cukup baik dalam membedakan kemampuan siswa. Dengan demikian, instrumen asesmen formatif yang dikembangkan dapat digunakan untuk mengukur kemampuan siswa pada materi Teorema Pythagoras.

Kata kunci: asesmen formatif, karakteristik psikometrik, kualitas instrumen, model Rasch

PENDAHULUAN

Asesmen merupakan bagian penting dalam proses pembelajaran yang berfungsi untuk memperoleh informasi tentang perkembangan dan pencapaian kompetensi siswa. Informasi yang diperoleh dapat digunakan guru untuk mengevaluasi efektivitas pembelajaran, mengidentifikasi kebutuhan belajar siswa dan merancang tindak lanjut yang sesuai (Badan Standar, Kurikulum, 2022). Dalam perkembangannya, asesmen tidak lagi dipandang hanya sebagai alat ukur untuk mengukur hasil belajar pada akhir pembelajaran, tetapi juga sebagai sarana yang mendukung proses belajar secara berkelanjutan dan membantu siswa mencapai tujuan pembelajaran yang diharapkan (Maskos et al., 2025);(Töllner & Kuhl, 2025).

Salah satu bentuk asesmen yang berperan penting dalam mendukung proses pembelajaran adalah asesmen formatif. Asesmen formatif merupakan proses pengumpulan dan penggunaan informasi selama proses pembelajaran berlangsung untuk memperoleh gambaran tentang perkembangan belajar siswa serta memberikan umpan balik yang dapat digunakan dalam memperbaiki proses pembelajaran (Töllner & Kuhl, 2025). Dalam pembelajaran matematika, asesmen formatif tidak hanya digunakan untuk memantau pencapaian kompetensi siswa, tetapi juga untuk mengidentifikasi kesulitan belajar, menyesuaikan strategi pembelajaran, dan mendukung keterlibatan siswa dalam proses belajar (Maskos et al., 2025). Selain itu, asesmen formatif memungkinkan guru memahami perkembangan dan kebutuhan belajar siswa sehingga pembelajaran dapat dirancang secara lebih terarah, relevan, dan berpusat pada siswa (Yin et al., 2022);(Pratama et al., 2024). Berbagai penelitian menunjukkan bahwa penerapan asesmen formatif memberikan dampak positif terhadap peningkatan hasil belajar, motivasi, partisipasi siswa, dan kemampuan berpikir matematis (Muhibuddin et al., 2024);(Maskos et al., 2025). Oleh karena itu, asesmen formatif menjadi salah satu bagian penting dalam pembelajaran matematika. Agar dapat memberikan informasi yang akurat dan bermakna bagi perbaikan pembelajaran, pelaksanaannya perlu didukung oleh instrumen asesmen yang berkualitas (Pastore, 2023);(Töllner & Kuhl, 2025).

Instrumen asesmen yang berkualitas merupakan prasyarat untuk menghasilkan informasi yang akurat tentang kemampuan dan perkembangan belajar siswa. Kualitas instrumen menentukan ketepatan data yang diperoleh sehingga berpengaruh terhadap tindak lanjut pembelajaran yang dilakukan oleh guru. Instrumen yang baik harus mampu mengukur karakteristik yang akan diukur secara konsisten sehingga menghasilkan data yang valid dan reliabel (Guo, 2024). Selain itu, pengujian kualitas instrumen perlu dilakukan untuk memastikan bahwa setiap butir soal dapat merepresentasikan kemampuan siswa secara akurat dan dapat membedakan siswa yang memiliki tingkat kemampuan yang berbeda (Moore et al., 2024). Dalam proses asesmen, validitas dan reliabilitas merupakan aspek penting yang menentukan kualitas instrumen sehingga hasil asesmen dapat menggambarkan kemampuan siswa secara akurat dan konsisten (American Educational Research Association; American Psychological Association; National Council on Measurement in Education, 2014). Oleh karena itu, analisis karakteristik psikometrik instrumen diperlukan untuk memastikan bahwa instrumen asesmen formatif dapat menghasilkan informasi yang akurat, valid dan reliabel sebagai dasar perbaikan pembelajaran.

Berbagai pendekatan telah digunakan untuk mengevaluasi kualitas instrumen asesmen, salah satunya adalah Teori Tes Klasik (*Classical Test Theory*). Pendekatan ini banyak digunakan karena relatif sederhana dalam menganalisis validitas, reliabilitas, tingkat kesukaran, dan daya pembeda butir soal (Rohmatdi et al., 2024). Namun, Teori Tes Klasik memiliki beberapa keterbatasan seperti karakteristik butir soal dan kemampuan siswa sangat dipengaruhi

oleh sampel yang digunakan sehingga hasil pengukuran cenderung kurang stabil ketika diterapkan pada kelompok yang berbeda (Nisfatulsanah et al., 2024). Seiring dengan perkembangan teori pengukuran modern, analisis kualitas instrumen saat ini banyak menggunakan pendekatan *Item Response Theory* (IRT) karena dapat mengestimasi kemampuan siswa dan karakteristik butir soal secara lebih objektif pada skala pengukuran yang sama (Chen & Science, 2021). Selain itu, pendekatan IRT memungkinkan analisis yang lebih mendalam terhadap kualitas instrumen, sehingga dapat memberikan informasi yang lebih lengkap terhadap karakteristik instrumen dibandingkan pengukuran klasik (Ulwiyah, 2023). Oleh karena itu, pendekatan modern *Item Response Theory* (IRT) menjadi alternatif yang banyak digunakan dalam pengembangan dan evaluasi instrumen untuk memperoleh hasil pengukuran yang lebih akurat dan konsisten.

Salah satu model yang banyak digunakan dalam pendekatan *Item Response Theory* (IRT) adalah Model Rasch. Model Rasch merupakan model pengukuran modern yang mengestimasi kemampuan siswa dan tingkat kesulitan butir soal secara bersamaan pada satu skala pengukuran yang sama. Berbeda dengan Teori Tes Klasik yang hasil pengukurannya sangat dipengaruhi oleh karakteristik sampel, Model Rasch memungkinkan parameter butir soal tetap relatif stabil meskipun diterapkan pada kelompok responden yang berbeda (Natsir et al., 2022). Selain itu, Model Rasch mampu memberikan informasi yang lebih detail tentang kualitas instrumen, seperti reliabilitas butir soal, tingkat kesulitan butir, kesesuaian butir soal (*item fit*), kesesuaian respons siswa (*person fit*), serta pola respons siswa terhadap instrumen yang digunakan (Bambang Sumintono, 2015). Oleh karena itu, Model Rasch menjadi salah satu pendekatan yang relevan untuk menganalisis karakteristik psikometrik instrumen asesmen formatif karena dapat memberikan informasi yang akurat, objektif, konsisten terhadap kualitas instrumen (Azisah, 2020);(Natsir et al., 2022).

Meskipun asesmen formatif telah banyak digunakan dalam pembelajaran matematika, kajian yang secara khusus menelaah kualitas psikometrik instrumen asesmen formatif pada materi tertentu masih perlu diperkuat. Sejumlah penelitian menunjukkan bahwa instrumen asesmen formatif perlu dianalisis secara lebih mendalam agar tidak hanya berfungsi sebagai alat pemberi skor, tetapi juga mampu menghasilkan informasi pengukuran yang valid, reliabel, dan bermakna bagi perbaikan pembelajaran. Dalam konteks ini, analisis berbasis Model Rasch semakin banyak digunakan karena mampu memberikan ukuran berbasis skala interval, menelaah fungsi butir, serta mengidentifikasi kualitas instrumen secara lebih rinci dibandingkan pendekatan pengukuran klasik. Yan dan Pastore (2022), misalnya, menggunakan analisis Rasch untuk menguji skala praktik asesmen formatif guru dan menunjukkan pentingnya bukti dimensionalitas, kualitas butir, serta kemungkinan adanya differential item functioning (DIF) pada konteks budaya yang berbeda. Beberapa penelitian telah memanfaatkan Model Rasch untuk mengevaluasi kualitas instrumen pembelajaran dan menunjukkan bahwa Model Rasch dapat digunakan untuk menganalisis berbagai karakteristik psikometrik instrumen, seperti reliabilitas, tingkat kesukaran butir soal, dan kesesuaian butir soal (Azisah, 2020);(Natsir et al., 2022).

Meskipun berbagai penelitian telah memanfaatkan Model Rasch untuk mengevaluasi kualitas instrumen pembelajaran, kajian yang secara khusus menganalisis karakteristik psikometrik instrumen asesmen formatif pada materi Teorema Pythagoras masih relatif terbatas sehingga diperlukan analisis lebih lanjut untuk memastikan kualitas instrumen yang digunakan. Oleh karena itu, penelitian ini bertujuan untuk menganalisis karakteristik psikometrik instrumen asesmen formatif pembelajaran matematika pada materi Teorema Pythagoras menggunakan

Model Rasch. Analisis instrumen difokuskan pada reliabilitas instrumen, tingkat kesukaran butir soal, kesesuaian butir soal (*item fit*), kemampuan siswa (*person measure*), serta pola respons siswa untuk memperoleh gambaran kualitas instrumen yang digunakan dalam pembelajaran matematika.

METODE

Pendekatan yang digunakan dalam penelitian ini adalah pendekatan kuantitatif yang bertujuan untuk menganalisis karakteristik psikometrik instrumen asesmen formatif menggunakan Model Rasch pada materi Teorema Pythagoras. Sampel penelitian terdiri atas 20 siswa kelas VIII SMP yang telah mempelajari materi Teorema Pythagoras yang dipilih sebagai responden untuk mengerjakan instrumen asesmen formatif yang dikembangkan. Instrumen penelitian berupa tes pilihan ganda sebanyak 20 butir soal pada materi Teorema Pythagoras.

Data penelitian diperoleh dari respons siswa terhadap instrumen yang diberikan setelah proses pembelajaran. Seluruh respons siswa kemudian dianalisis menggunakan perangkat Winsteps (Linacre, 2009) yang bertujuan untuk mengevaluasi karakteristik psikometrik instrumen yang meliputi reliabilitas instrumen (*Cronbach Alpha*), *person reliability*, *item reliability*, tingkat kesukaran butir soal, kesesuaian butir soal (*item fit*), kemampuan siswa (*person measure*), serta pola respons siswa. Kesesuaian butir soal dianalisis berdasarkan nilai Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), dan Point Measure Correlation (Pt Measure Corr). Nilai Outfit MNSQ adalah $0,5 < \text{MNSQ} < 1,5$; Outfit Nilai ZSTD berkisar dari $-2,0 < \text{ZSTD} < 2,0$; dan nilai korelasi dari skor total (*Point Measure Correlation*) berkisar dari $0,4 < \text{Pt M Corr} < 0,85$ (Bambang Sumintono, 2015); (Boone, W.J., Staver, J.R., & Yale, M.S, 2014).

Kriteria dalam menentukan nilai Item Reliabilitas dan Person Reliabilitas dalam Model Rasch didasarkan pada Table 1.

Tabel. Kriteria Interpretasi Nilai Item Reliability dan Person Reliability dalam Model Rasch

No.	Nilai Reliabilitas	Kategori
1	$> 0,94$	Istimewa
2	$0,91 - 0,94$	Sangat baik
3	$0,81 - 0,90$	Baik
4	$0,67 - 0,80$	Cukup
5	$< 0,67$	Lemah

Sumber: Sumintono (2015).

Tingkat kesukaran butir soal ditentukan berdasarkan nilai logit item yang dihasilkan dari analisis Model Rasch. Tingkat kesukaran butir soal (*item measure*) ditentukan berdasarkan nilai logit pada kolom *measure*. Semakin tinggi nilai logit suatu butir soal, semakin tinggi pula tingkat kesukaran butir soal tersebut. Sebaliknya, semakin rendah nilai logit, semakin mudah butir soal tersebut bagi responden. Pengelompokan tingkat kesukaran dilakukan berdasarkan nilai rata-rata logit item dan standar deviasi (SD) sehingga butir soal dapat dikategorikan menjadi sangat sulit, sulit, mudah, dan sangat mudah (Sumintono, 2015). Pada Tabel 2 berikut disajikan kriteria tingkat kesukaran butir soal berdasarkan nilai measure logit dalam bentuk tabel:

Tabel 2. Kriteria Kategorisasi Tingkat Kesukaran Butir Soal Berdasarkan Nilai Measure Logit

No.	Kriteria Nilai Measure Logit	Kategori Tingkat Kesukaran
1	Measure logit $< - \text{SD logit}$	Item sangat mudah
2	$- \text{SD logit} \leq \text{Measure logit} \leq 0$	Item mudah

3	$0 \leq \text{Measure logit} \leq \text{SD logit}$	Item sulit
4	$\text{Measure logit} > \text{SD logit}$	Item sangat sulit

Sumber: Sumintono (2015).

Selanjutnya, Wright Map digunakan untuk memetakan hubungan antara kemampuan siswa dan tingkat kesukaran butir soal sehingga diperoleh gambaran distribusi kemampuan siswa dan sebaran tingkat kesukaran soal. Selain itu, analisis Scalogram digunakan untuk mengidentifikasi pola respons siswa serta mengungkap adanya respons yang tidak konsisten. Hasil analisis diinterpretasikan secara deskriptif untuk memperoleh gambaran kualitas instrumen asesmen formatif pada materi Teorema Pythagoras berdasarkan karakteristik psikometrik yang dihasilkan oleh Model Rasch.

HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk menganalisis karakteristik psikometrik instrumen asesmen formatif pada materi Teorema Pythagoras menggunakan Model Rasch. Analisis dilakukan untuk memperoleh informasi mengenai kualitas instrumen berdasarkan reliabilitas, kesesuaian butir soal terhadap model, tingkat kesukaran butir, kemampuan siswa, serta pola respons siswa. Melalui analisis ini, instrumen tidak hanya dinilai berdasarkan skor akhir siswa, tetapi juga ditelaah berdasarkan fungsi setiap butir soal dalam mengukur kemampuan siswa secara lebih objektif. Hasil analisis disajikan dan dibahas secara bertahap sebagai berikut.

1. Reliabilitas Instrumen

Hasil analisis reliabilitas instrumen tes kemampuan siswa pada materi Teorema Pythagoras dapat dilihat pada Tabel 1.

Tabel 3. Hasil Analisis Reliabilitas Instrumen

	Reliabilitas
Person	0,67
Item	0,77

Berdasarkan Tabel 3, hasil analisis reliabilitas menunjukkan bahwa instrumen memiliki tingkat reliabilitas yang cukup baik untuk digunakan dalam mengukur kemampuan siswa. Nilai *item reliability* yang lebih tinggi dibandingkan *person reliability* menunjukkan bahwa karakteristik tingkat kesukaran butir soal relatif lebih konsisten dibandingkan konsistensi respons siswa dalam mengerjakan tes. Kondisi ini dapat disebabkan oleh adanya variasi kemampuan antarsiswa yang menghasilkan pola respons yang beragam. Nilai *item reliability* sebesar 0,77 termasuk dalam kategori cukup, yang menunjukkan bahwa butir soal telah memiliki tingkat kesukaran yang tersusun relatif konsisten sehingga mampu membedakan kemampuan siswa, meskipun kualitas instrumen masih dapat ditingkatkan melalui penyempurnaan butir soal dan penambahan jumlah responden (Bambang Sumintono, 2015). Hasil penelitian ini, sejalan dengan penelitian (Boone, W.J., Staver, J.R., & Yale, M.S, 2014) yang menyatakan bahwa nilai reliabilitas item yang semakin tinggi menunjukkan parameter tingkat kesukaran butir soal lebih konsisten sehingga hasil pengukuran yang diperoleh lebih akurat.

2. Kesesuaian Butir Soal terhadap Model Rasch

Kesesuaian butir soal terhadap Model Rasch dianalisis untuk mengetahui sejauh mana setiap butir berfungsi sesuai dengan konstruk kemampuan yang diukur. Analisis ini penting

dilakukan karena butir soal yang tidak sesuai dengan model dapat menunjukkan adanya masalah, seperti redaksi soal yang kurang jelas, kunci jawaban yang tidak tepat, distraktor yang tidak berfungsi, atau ketidaksesuaian antara butir soal dengan indikator kemampuan yang diukur. Dalam penelitian ini, kesesuaian butir dianalisis berdasarkan nilai Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), dan Point Measure Correlation (Pt Measure Corr). Butir soal dinyatakan sesuai dengan Model Rasch apabila nilai statistiknya berada pada rentang kriteria yang telah ditetapkan. Hasil analisis kesesuaian butir soal disajikan pada Tabel 4 berikut.

Tabel 4. Item Statistics: Misfit Order

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ ZSTD	OUTFIT MNSQ ZSTD	PTMEASUR-AL CORR. EXP.	EXACT MATCH OBS% EXP%	Item
1	7	20	1.34	.52	1.52 2.20	1.86 2.22	A-.11 .41	60.0 70.5	1
16	5	20	1.91	.56	1.26 1.01	1.50 1.07	B .09 .37	75.0 75.7	16
15	2	20	3.13	.77	1.11 .37	1.41 .70	C .09 .25	90.0 89.9	15
5	10	20	.58	.50	1.30 1.38	1.38 1.48	D .17 .44	60.0 70.7	5
19	13	20	-.18	.52	1.26 1.16	1.23 .79	E .20 .43	65.0 71.7	19
9	9	20	.83	.50	1.23 1.12	1.20 .83	F .25 .44	55.0 70.4	9
13	16	20	-1.09	.60	1.12 .46	.94 .08	G .28 .35	80.0 79.9	13
4	18	20	-1.99	.77	1.05 .27	.72 .00	H .26 .26	90.0 89.9	4
17	13	20	-.18	.52	1.03 .23	.92 -.15	I .42 .43	65.0 71.7	17
14	14	20	-.46	.54	1.01 .14	1.01 .16	J .40 .41	75.0 73.3	14
18	13	20	-.18	.52	.96 -.11	.94 -.09	i .46 .43	75.0 71.7	18
3	15	20	-.76	.56	.88 -.38	.75 -.45	h .51 .38	85.0 76.2	3
8	12	20	.08	.51	.83 -.75	.80 -.67	g .59 .44	75.0 70.4	8
7	9	20	.83	.50	.77 -1.12	.78 -.85	f .63 .44	85.0 70.4	7
10	15	20	-.76	.56	.78 -.81	.59 -.87	e .60 .38	85.0 76.2	10
20	14	20	-.46	.54	.75 -1.07	.58 -1.16	d .66 .41	75.0 73.3	20
6	16	20	-1.09	.60	.72 -.88	.50 -.90	c .62 .35	80.0 79.9	6
12	15	20	-.76	.56	.72 -1.09	.52 -1.13	b .66 .38	75.0 76.2	12
2	15	20	-.76	.56	.67 -1.32	.49 -1.20	a .70 .38	85.0 76.2	2
MEAN	12.5	20.0	-.20	.63	1.00 .04	.95 -.01		75.5 75.5	
P.SD	4.3	.0	1.44	.29	.24 .98	.38 .95		10.1 5.8	

Hasil analisis kesesuaian butir soal menggunakan Model Rasch, sebagian besar butir soal telah memenuhi kriteria kesesuaian model. Analisis dilakukan berdasarkan nilai Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), dan Point Measure Correlation (Pt Measure Corr). Hasil analisis menunjukkan bahwa sebagian besar butir soal memiliki nilai Outfit MNSQ pada rentang 0,5–1,5, nilai Outfit ZSTD pada rentang -2,0 hingga +2,0, serta nilai Point Measure Correlation yang positif. Namun, terdapat satu butir soal yang menunjukkan ketidaksesuaian dengan Model Rasch, yaitu butir soal nomor 1. Butir tersebut memiliki nilai Outfit MNSQ sebesar 1,86, Outfit ZSTD sebesar 2,22, dan Point Measure Correlation sebesar -0,11. Nilai tersebut berada di luar rentang yang direkomendasikan sehingga menunjukkan bahwa butir soal belum berfungsi sebagaimana yang diharapkan dalam mengukur kemampuan siswa. Oleh karena itu, butir soal nomor 1 perlu direvisi sebelum digunakan kembali dalam kegiatan asesmen. Berdasarkan hasil analisis dapat disimpulkan bahwa sebagian besar butir soal telah sesuai dengan Model Rasch sehingga instrumen layak digunakan untuk mengukur kemampuan siswa pada materi Teorema Pythagoras.

3. Tingkat Kesukaran Butir Soal

Tingkat kesukaran butir soal dianalisis berdasarkan nilai item measure yang dihasilkan melalui Model Rasch. Nilai item measure dinyatakan dalam satuan logit, yang menunjukkan posisi tingkat kesukaran setiap butir soal pada skala pengukuran yang sama. Semakin tinggi nilai logit suatu butir, semakin tinggi tingkat kesukaran butir tersebut. Sebaliknya, semakin rendah nilai logit suatu butir, semakin mudah butir tersebut dijawab oleh siswa.

Analisis tingkat kesukaran butir penting dilakukan untuk mengetahui apakah instrumen asesmen formatif memiliki sebaran soal yang bervariasi, mulai dari soal sangat mudah, mudah, sulit, hingga sangat sulit. Sebaran tingkat kesukaran yang baik memungkinkan instrumen mengukur kemampuan siswa pada berbagai level. Dengan demikian, guru dapat memperoleh informasi yang lebih rinci mengenai kemampuan siswa serta menentukan tindak lanjut pembelajaran yang sesuai. Hasil analisis tingkat kesukaran butir soal berdasarkan output Model Rasch disajikan pada Tabel 5.

Tabel 5. Item Statistics: Measure Order

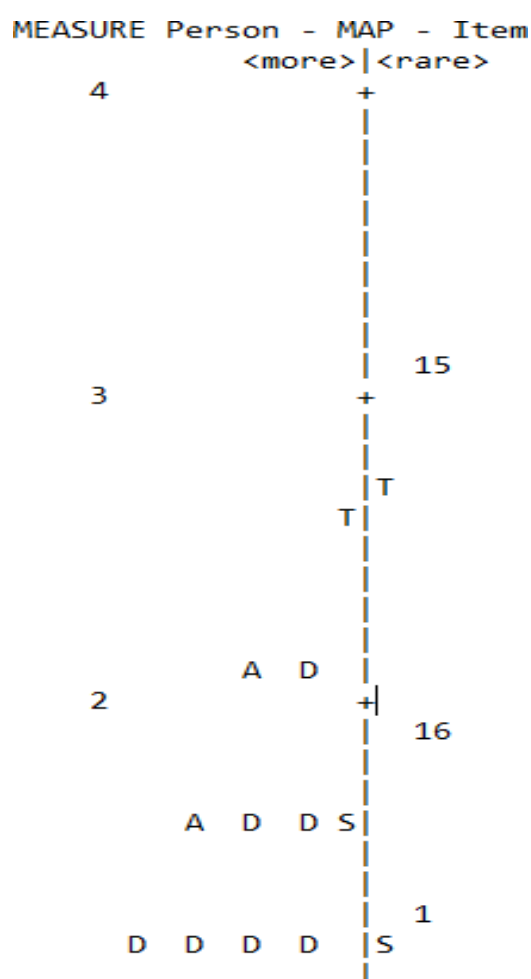
ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ ZSTD	OUTFIT MNSQ ZSTD	PTMEASUR-AL CORR. EXP.	EXACT MATCH OBS% EXP%	Item
15	2	20	3.13	.77	1.11 .37	1.41 .70	.09 .25	90.0 89.9	15
16	5	20	1.91	.56	1.26 1.01	1.50 1.07	.09 .37	75.0 75.7	16
1	7	20	1.34	.52	1.52 2.20	1.86 2.22	-.11 .41	60.0 70.5	1
7	9	20	.83	.50	.77 -1.12	.78 -.85	.63 .44	85.0 70.4	7
9	9	20	.83	.50	1.23 1.12	1.20 .83	.25 .44	55.0 70.4	9
5	10	20	.58	.50	1.30 1.38	1.38 1.48	.17 .44	60.0 70.7	5
8	12	20	.08	.51	.83 -.75	.80 -.67	.59 .44	75.0 70.4	8
17	13	20	-.18	.52	1.03 .23	.92 -.15	.42 .43	65.0 71.7	17
18	13	20	-.18	.52	.96 -.11	.94 -.09	.46 .43	75.0 71.7	18
19	13	20	-.18	.52	1.26 1.16	1.23 .79	.20 .43	65.0 71.7	19
14	14	20	-.46	.54	1.01 .14	1.01 .16	.40 .41	75.0 73.3	14
20	14	20	-.46	.54	.75 -1.07	.58 -1.16	.66 .41	75.0 73.3	20
2	15	20	-.76	.56	.67 -1.32	.49 -1.20	.70 .38	85.0 76.2	2
3	15	20	-.76	.56	.88 -.38	.75 -.45	.51 .38	85.0 76.2	3
10	15	20	-.76	.56	.78 -.81	.59 -.87	.60 .38	85.0 76.2	10
12	15	20	-.76	.56	.72 -1.09	.52 -1.13	.66 .38	75.0 76.2	12
6	16	20	-1.09	.60	.72 -.88	.50 -.90	.62 .35	80.0 79.9	6
13	16	20	-1.09	.60	1.12 .46	.94 .08	.28 .35	80.0 79.9	13
4	18	20	-1.99	.77	1.05 .27	.72 .00	.26 .26	90.0 89.9	4
11	20	20	-4.05	1.84	MINIMUM MEASURE		.00 .00	100.0 100.0	11
MEAN	12.5	20.0	-.20	.63	1.00 .04	.95 -.01		75.5 75.5	
P.SD	4.3	.0	1.44	.29	.24 .98	.38 .95		10.1 5.8	

Berdasarkan nilai rata-rata (*mean*) dan standar deviasi, terdapat 3 butir soal yang termasuk dalam kategori sangat sulit yaitu soal nomor 1, 15 dan 16. Selanjutnya, terdapat 8 butir soal yang termasuk dalam kategori sulit, yaitu butir soal nomor 5, 7, 8, 9, 14, 17, 18, dan 19. Sebanyak 7 butir soal yang termasuk dalam kategori mudah, yaitu 2, 3, 6, 10, 12, 13, dan 20. Dan terdapat 2 butir soal yang termasuk dalam kategori sangat mudah, yaitu butir nomor 4 dan 11. Hasil tersebut menunjukkan bahwa instrumen memiliki variasi tingkat kesukaran yang cukup beragam sehingga dapat mengukur kemampuan siswa pada berbagai tingkat kemampuan. Hasil penelitian ini sejalan dengan penelitian (Azizah et al., 2022);(Uba, 2024);(Boone, 2016);(Hamdu et al., 2020) yang menyatakan bahwa variasi tingkat kesukaran butir soal

diperlukan agar instrumen dapat mengukur siswa pada berbagai tingkat kemampuan dan menghasilkan informasi pengukuran yang lebih akurat.

4. Kemampuan Siswa

Setelah tingkat kesukaran butir soal dianalisis, langkah berikutnya adalah memetakan hubungan antara kemampuan siswa dan tingkat kesukaran butir soal dalam satu skala pengukuran. Pemetaan ini dilakukan melalui Wright Map, yaitu peta yang menampilkan posisi kemampuan siswa pada sisi kiri dan tingkat kesukaran butir soal pada sisi kanan. Melalui Wright Map, dapat diketahui apakah sebaran butir soal telah sesuai dengan variasi kemampuan siswa. Selain itu, Wright Map juga membantu mengidentifikasi butir yang terlalu mudah, terlalu sulit, maupun butir yang berada pada tingkat kesukaran yang seimbang dengan kemampuan responden. Hasil pemetaan kemampuan siswa dan tingkat kesukaran butir soal disajikan pada Gambar 1.



Gambar 1. Wright Map

Berdasarkan Wright Map, butir soal nomor 15 memiliki tingkat kesukaran tertinggi. Selanjutnya, butir soal nomor 16 dan nomor 1 juga menunjukkan tingkat kesukaran yang relatif tinggi dibandingkan butir soal lainnya. Hanya sebagian kecil siswa yang memiliki kemampuan untuk menjawab soal nomor 15, 16 dan 1. Hal ini menunjukkan bahwa ketiga butir soal tersebut merupakan soal yang paling menantang bagi siswa karena tidak semua siswa dapat menjawab

dengan benar. Sedangkan, butir soal nomor 11 dan nomor 4 merupakan soal yang paling mudah karena dapat dijawab dengan benar oleh seluruh siswa.

5. Pola Respons Siswa

Selain memetakan hubungan antara kemampuan siswa dan tingkat kesukaran butir soal melalui Wright Map, analisis juga dilakukan terhadap pola respons siswa menggunakan Scalogram. Analisis Scalogram bertujuan untuk melihat keteraturan jawaban siswa berdasarkan urutan kemampuan responden dan tingkat kesukaran butir soal. Melalui analisis ini, dapat diketahui apakah siswa dengan kemampuan tinggi cenderung menjawab lebih banyak soal dengan benar, serta apakah terdapat pola respons yang tidak konsisten, seperti siswa yang mampu menjawab soal sulit tetapi salah pada soal yang lebih mudah.

Pola respons yang tidak konsisten penting diperhatikan karena dapat mengindikasikan adanya kemungkinan siswa menebak jawaban, kurang teliti, mengalami miskonsepsi, atau menghadapi kendala dalam memahami redaksi soal tertentu. Oleh karena itu, Scalogram tidak hanya memberikan informasi tentang performa siswa, tetapi juga dapat menjadi dasar untuk meninjau kembali kualitas butir soal dan keakuratan respons siswa. Hasil analisis pola respons siswa melalui Scalogram disajikan pada Gambar 2.

GUTTMAN SCALOGRAM OF RESPONSES :

Person	Item	
	1 1 1112111 11	
	14632302407898579165	
8	+11111111111111111000	D
9	+1111111111111111001110	A
10	+111111110111110111001	D
11	+111111111111111110000	D
17	+111111111111011010110	A
5	+1111111111110111011000	D
13	+111111111111101110000	D
14	+111111111111101110000	D
15	+111111111111101110000	D
16	+111010111111111001100	A
2	+1111111011110010001010	D
1	+101101111110000101010	D
7	+110111011011110001000	D
19	+11101011000111000101	A
4	+11111110000001100100	A
3	+11101110001100100000	D
12	+10100010010011010110	A
18	+11010100011010001000	D
6	+11010001100110000000	D
20	+11010000101000100100	A
	1 1 1112111 11	
	14632302407898579165	

Gambar 2. Scalogram

Berdasarkan hasil analisis Scalogram pada Gambar 2, sebagian besar siswa menunjukkan pola respons yang sesuai dengan tingkat kemampuannya. Siswa dengan kemampuan tinggi cenderung dapat menjawab lebih banyak soal dengan benar dibandingkan dengan siswa yang memiliki kemampuan lebih rendah. Namun, terdapat beberapa siswa yang menunjukkan pola respons yang tidak konsisten. Seperti, siswa nomor 19 mampu menjawab beberapa soal yang tergolong sulit, tetapi tidak dapat menjawab beberapa soal yang tergolong mudah. hal ini dapat terjadi karena siswa hanya menebak jawaban, kurang teliti saat mengerjakan soal, mengalami kesulitan dalam memahami soal tertentu. Meskipun demikian, secara menyeluruh pola respons

siswa menunjukkan bahwa data yang diperoleh cukup konsisten dan dapat digunakan untuk analisis lebih lanjut. Menurut (Meijer, R.R., & Sitsma, 2001), analisis *person fit* dapat digunakan untuk mengidentifikasi pola respons siswa yang tidak biasa. Apabila pola respons menunjukkan tingkat ketidakpastian yang tinggi, hal tersebut dapat mengindikasikan bahwa siswa mengalami kendala atau memberikan respons yang tidak konsisten selama mengerjakan tes (Boone, W.J., Staver, J.R., & Yale, M.S, 2014);(Susongko et al., 2019) .

Hasil analisis Rasch dalam penelitian ini memberikan informasi penting bagi penguatan fungsi asesmen formatif dalam pembelajaran matematika. Asesmen formatif tidak hanya berperan untuk mengetahui capaian siswa, tetapi juga menjadi dasar bagi guru dalam memberikan umpan balik, mengidentifikasi kesulitan belajar, serta merancang tindak lanjut pembelajaran yang sesuai. Dalam konteks ini, Model Rasch memiliki keunggulan karena mampu menghubungkan kemampuan siswa dan tingkat kesukaran butir soal pada satu skala logit yang sama. Dengan demikian, hasil pengukuran tidak hanya menunjukkan skor benar atau salah, tetapi juga memberikan informasi diagnostik mengenai posisi kemampuan siswa terhadap tingkat kesukaran soal. Hal ini sejalan dengan pandangan Lestari dan Samsudin (2020), Hamdu et al. (2020), serta Sumintono (2018) bahwa analisis Rasch dapat memperkuat asesmen formatif karena menghasilkan informasi pengukuran yang lebih presisi, berbasis bukti, dan dapat digunakan untuk memantau perkembangan belajar siswa.

Nilai reliabilitas, person reliability, dan item reliability dalam penelitian ini memberikan gambaran mengenai ketepatan instrumen dalam mengukur kemampuan siswa. Person reliability menunjukkan konsistensi pola kemampuan siswa, sedangkan item reliability menunjukkan kestabilan hierarki tingkat kesukaran butir soal. Dalam asesmen formatif, kedua informasi ini penting karena guru membutuhkan instrumen yang tidak hanya reliabel secara umum, tetapi juga mampu membedakan siswa berdasarkan tingkat kemampuan dan memantau perkembangan mereka dari waktu ke waktu. Hamdu et al. (2020) dan Sumintono (2018) menegaskan bahwa reliabilitas dan separation dalam analisis Rasch dapat digunakan untuk melihat ketepatan instrumen dalam membedakan kemampuan responden dan kualitas butir. Oleh karena itu, reliabilitas yang berada pada kategori cukup dalam penelitian ini menunjukkan bahwa instrumen telah memiliki dasar pengukuran yang dapat digunakan, tetapi masih perlu diperkuat melalui revisi butir dan pengujian pada jumlah responden yang lebih besar.

Selain reliabilitas, analisis item fit juga menjadi bagian penting dalam menilai kualitas instrumen asesmen formatif. Item fit melalui nilai Outfit MNSQ, ZSTD, dan Point Measure Correlation digunakan untuk melihat apakah setiap butir soal berfungsi sesuai dengan konstruk kemampuan yang diukur. Butir yang tidak sesuai dengan Model Rasch dapat mengganggu ketepatan informasi asesmen karena berpotensi menghasilkan umpan balik yang keliru kepada siswa. Widodo dan Chotimah (2023), Aprilia et al. (2021), dan Hamdu et al. (2020) menjelaskan bahwa item fit dapat digunakan untuk mendeteksi butir yang tidak bekerja secara optimal, misalnya karena redaksi soal kurang jelas, distraktor tidak efektif, kunci jawaban bermasalah, atau indikator soal tidak sesuai dengan kemampuan yang hendak diukur. Dengan demikian, butir yang teridentifikasi misfit dalam penelitian ini perlu direvisi agar tidak mengaburkan interpretasi kemampuan siswa dan tidak mengganggu fungsi asesmen formatif sebagai dasar pemberian umpan balik.

Dalam konteks materi Teorema Pythagoras, hasil analisis tingkat kesukaran butir memberikan informasi pedagogis yang sangat penting. Butir yang tergolong sangat mudah dapat menunjukkan bahwa siswa telah menguasai konsep dasar, seperti mengenali unsur segitiga siku-siku atau menggunakan rumus secara langsung. Sebaliknya, butir yang tergolong sulit atau

sangat sulit kemungkinan menuntut kemampuan berpikir yang lebih kompleks, seperti memahami hubungan antarsisi segitiga, menafsirkan gambar, memilih strategi penyelesaian, serta menerapkan konsep Teorema Pythagoras dalam masalah kontekstual. Lestari dan Samsudin (2020), Apriani et al. (2023), serta Setyawati et al. (2023) menegaskan bahwa item measure dalam Rasch dapat membantu guru mengidentifikasi konsep yang masih sulit dipahami siswa sehingga pembelajaran dapat diarahkan pada latihan yang lebih tepat, penjelasan bertahap, dan pemeriksaan pemahaman secara berkala. Oleh karena itu, informasi tingkat kesukaran butir tidak hanya berfungsi untuk menilai kualitas soal, tetapi juga untuk menentukan bagian materi yang perlu diperkuat dalam pembelajaran.

Wright Map dalam penelitian ini juga memberikan kontribusi penting bagi asesmen formatif karena memetakan kemampuan siswa dan tingkat kesukaran butir dalam satu skala logit. Peta ini membantu guru mengetahui apakah sebaran soal telah sesuai dengan variasi kemampuan siswa. Jika terdapat butir yang terlalu mudah bagi sebagian besar siswa, maka butir tersebut mungkin kurang mampu membedakan kemampuan siswa pada level menengah dan tinggi. Sebaliknya, jika terdapat butir yang terlalu sulit, guru perlu meninjau apakah kesulitan tersebut muncul karena tuntutan kognitif soal memang tinggi atau karena siswa belum memperoleh pemahaman yang cukup terhadap konsep yang diukur. Almubarak et al. (2023), Wijaya (2021), dan Aprilia et al. (2021) menjelaskan bahwa Wright Map dapat digunakan untuk mengidentifikasi kesenjangan antara kemampuan siswa dan tingkat kesukaran soal, sehingga guru dapat merancang scaffolding, pengayaan, atau pengulangan materi secara lebih terarah. Dengan demikian, Wright Map dapat membantu guru menyesuaikan urutan pembelajaran dengan kebutuhan aktual siswa.

Selain Wright Map, Scalogram memberikan informasi mengenai pola respons siswa dalam menjawab butir soal. Pola respons yang tidak konsisten, misalnya siswa menjawab benar pada soal sulit tetapi salah pada soal yang lebih mudah, dapat mengindikasikan beberapa kemungkinan, seperti menebak jawaban, kurang teliti, miskonsepsi, atau kesulitan memahami redaksi soal. Apriani et al. (2023), Diyanah dan Yusmaita (2022), serta Lestari dan Samsudin (2020) menyatakan bahwa Scalogram dapat digunakan untuk mengungkap pola respons yang tidak biasa sehingga guru dapat memberikan umpan balik langsung dan melakukan penyesuaian pembelajaran. Dalam asesmen formatif, informasi ini sangat penting karena guru tidak hanya melihat jumlah jawaban benar, tetapi juga dapat menelusuri bagaimana konsistensi siswa dalam menjawab soal berdasarkan tingkat kesukarannya.

Penggunaan Wright Map dan Scalogram secara bersamaan dapat memperkuat fungsi diagnostik asesmen formatif. Wright Map menunjukkan hubungan antara kemampuan siswa dan tingkat kesukaran soal, sedangkan Scalogram menunjukkan keteraturan atau ketidakteraturan pola respons siswa. Apriani dan Susilawati (2023), Setyawati et al. (2023), dan Risdianto et al. (2021) menegaskan bahwa kombinasi informasi dari Wright Map dan Scalogram dapat membantu guru menyesuaikan umpan balik, menyusun ulang urutan pembelajaran, serta mengurangi ketidaksesuaian antara kemampuan siswa dan tingkat kesukaran butir soal. Dengan demikian, hasil analisis Rasch tidak hanya bermanfaat untuk menilai kualitas instrumen, tetapi juga dapat digunakan untuk memperbaiki alur pembelajaran, menyusun latihan bertahap, dan merancang strategi pengajaran yang lebih responsif terhadap kebutuhan siswa.

Temuan penelitian ini juga sejalan dengan penelitian terdahulu yang menunjukkan bahwa Model Rasch dapat digunakan untuk merevisi dan mengembangkan instrumen asesmen matematika secara lebih valid dan reliabel. Hamdu et al. (2020) dan Sumintono (2018) menekankan pentingnya reliabilitas dan separation untuk memastikan ketepatan pengukuran.

Widodo dan Chotimah (2023), Aprilia et al. (2021), dan Hamdu et al. (2020) menegaskan bahwa item fit perlu dianalisis untuk mendeteksi butir yang berpotensi menyimpang dari konstruk pengukuran. Almubarak et al. (2023), Wijaya (2021), dan Aprilia et al. (2021) menunjukkan bahwa Wright Map dapat digunakan untuk mengevaluasi hierarki tingkat kesukaran butir dan kesesuaiannya dengan kemampuan responden. Sementara itu, Apriani et al. (2023), Diyanah dan Yusmaita (2022), serta Lestari dan Samsudin (2020) menunjukkan bahwa Scalogram dapat digunakan untuk membaca pola respons siswa dan mendeteksi kemungkinan tebakan atau ketidaktelitian. Dengan demikian, penelitian ini memperkuat temuan sebelumnya bahwa analisis Rasch memberikan informasi yang lebih mendalam dibandingkan analisis skor biasa.

Meskipun demikian, penelitian ini memiliki beberapa keterbatasan. Pertama, jumlah responden masih terbatas sehingga hasil penelitian perlu dipahami sebagai analisis awal terhadap kualitas instrumen. Jumlah responden yang kecil dapat memengaruhi kestabilan estimasi parameter item dan person. Kedua, cakupan materi hanya terbatas pada Teorema Pythagoras, sehingga hasil penelitian belum dapat digeneralisasi pada materi matematika lainnya. Ketiga, instrumen berbentuk pilihan ganda sehingga belum sepenuhnya mampu menangkap proses berpikir, strategi penyelesaian, atau alasan siswa dalam memilih jawaban. Keempat, penelitian ini belum mencakup analisis lanjutan seperti unidimensionalitas, differential item functioning, dan person fit secara lebih mendalam. Oleh karena itu, interpretasi hasil perlu dilakukan secara hati-hati dan perlu diperkuat melalui penelitian lanjutan dengan jumlah responden yang lebih besar serta cakupan analisis yang lebih luas.

Berdasarkan hasil dan keterbatasan tersebut, pengembangan instrumen asesmen formatif pada penelitian selanjutnya perlu diarahkan pada beberapa langkah. Pertama, butir yang misfit perlu direvisi dengan meninjau kembali redaksi soal, kesesuaian indikator, kunci jawaban, dan kualitas distraktor. Kedua, butir yang terlalu mudah atau terlalu sulit perlu dikalibrasi agar sebaran tingkat kesukaran lebih seimbang dengan kemampuan siswa. Ketiga, instrumen perlu diujicobakan pada jumlah responden yang lebih besar dan konteks sekolah yang lebih beragam agar estimasi parameter menjadi lebih stabil. Keempat, analisis lanjutan seperti Wright Map, Scalogram, unidimensionalitas, DIF, dan person fit perlu digunakan untuk memperkuat bukti validitas, reliabilitas, dan fungsi butir. Dengan demikian, instrumen asesmen formatif pada materi Teorema Pythagoras dapat dikembangkan secara lebih akurat, adil, dan bermanfaat untuk mendukung pembelajaran matematika yang adaptif dan berkelanjutan.

SIMPULAN DAN SARAN

Berdasarkan analisis Model Rasch, instrumen asesmen formatif pada materi Teorema Pythagoras memiliki karakteristik psikometrik yang cukup baik, yang ditunjukkan oleh reliabilitas instrumen yang berada pada kategori cukup, sebagian besar butir soal sesuai dengan model, serta variasi tingkat kesukaran yang mampu mengukur kemampuan siswa pada berbagai tingkat kemampuan. Dengan demikian, instrumen yang dikembangkan layak digunakan sebagai alat asesmen formatif dalam pembelajaran matematika. Namun, butir soal yang belum memenuhi kriteria kesesuaian perlu direvisi, dan penelitian selanjutnya disarankan melibatkan jumlah responden yang lebih besar untuk memperoleh hasil pengukuran yang lebih konsisten dan akurat.

DAFTAR PUSTAKA

- Almubarak, A., Saadi, P., Prayogi, R., & Maldini, P. P. (2023). Assessing Students Understanding of Chemical Bonds Material by Rasch Modeling. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, 217–232. <https://doi.org/10.23917/ijolae.v5i3.22242>
- American Educational Research Association; American Psychological Association; National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.
- Apriani, I. F., & Susilawati, L. (2023). Kemampuan Perkalian dan Pembagian Bilangan Cacah Menggunakan RASCH Model untuk Siswa Kelas IV SD. *Dwija Cendekia Jurnal Riset Pedagogik*, 7(2). <https://doi.org/10.20961/jdc.v7i2.75625>
- Apriani, I. F., Islamiati, G., Putri, I. R., Susilawati, L., Muharram, M. R. W., & Nur'aeni, E. (2023). Penggunaan model RASCH untuk soal perkalian bilangan cacah Siswa Sekolah Dasar. *Collase (Creative of Learning Students Elementary Education)*, 6(1), 86–93. <https://doi.org/10.22460/collase.v1i1.14018>
- Aprilia, N., Susilaningih, E., Sudarmin, S., Sumarni, W., Mahatmanti, F. Widhi, & Adhelia, N. U. (2021). Application Of The Rasch Model On The Test Instrument To Analyze The Problem-Solving Ability Of Students In Acid-Base Solutions Materials. *Edusains*, 13(2), 106–118. <https://doi.org/10.15408/es.v13i2.21375>
- Azisah, S. W. (2020). Model, Penggunaan Pada, Rasch Untuk Analisis Instrumen Tes Aktuaria, Matakuliah Matematika. *Jupitek*, 3, 45–50. [https://doi.org/https://doi.org/10.30598/jupitekvol3iss1ppx45-](https://doi.org/10.30598/jupitekvol3iss1ppx45-)
- Azizah, N., Suseno, M., & Hayat, B. (2022). *Item Analysis of the Rasch Model Items in the Final Semester Exam Indonesian Language Lesson*. 12(1), 15–26. <https://doi.org/10.5430/wjel.v12n1p15>
- Badan Standar, Kurikulum, dan A. P. (2022). *Panduan Pembelajaran dan Asesmen*.
- Bambang Sumintono, W. (2015). *Rasch*. Trim Komunika.
- Boone, W.J., Staver, J.R., & Yale, M.S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer.
- Boone, W. J. (2016). *Rasch Analysis for Instrument Development : Why , When , and How ?* 1–7. <https://doi.org/10.1187/cbe.16-04-0148>
- Chen, Y., & Science, P. (2021). *Item Response Theory – A Statistical Framework for Educational and Psychological Measurement*. 1–64.
- Diyanah, D., & Yusmaita, E. (2022). Analysis of SMA Pembangunan Laboratorium UNP Students' Chemical Literacy Ability on Reaction Rate Topic by Using Rasch Model. *Entalpi Pendidikan Kimia*, 3(1), 26–39. <https://doi.org/10.24036/epk.v3i1.245>
- Guo, S. (2024). *Validating an Instrument for Teachers ' Acceptance of Artificial Intelligence in Education*.
- Hamdu, G., Fuadi, F. N., Yulianto, A., & Akhirani, Y. S. (2020). *Items Quality Analysis Using Rasch Model To Measure Elementary School Students ' Critical Thinking Skill On Stem Learning*. 9(1), 61–74. <https://doi.org/10.23887/jpi-undiksha.v9i1.20884>

- Lestari, A., & Samsudin, A. (2020). Using Rasch Model Analysis to Analyze Students' Scientific Literacy on Heat and Temperature. <https://doi.org/10.4108/eai.12-10-2019.2296483>
- Linacre, J. . (2009). *A user's guide to WINSTEPS*. IL: Winsteps.
- Maskos, K., Schulz, A., Oeksuez, S. S., & Rakoczy, K. (2025). *Formative Assessment in Mathematics Education : A Systematic Review*. 679–693. <https://doi.org/10.1007/s11858-025-01696-x>
- Meijer, R.R., & Sitsma, K. (2001). *Person fit statistic: What is their purpose?. Rasch Measurement Transactions*. 15.
- Moore, S., Costello, E., & Huy, A. (2024). *An Automatic Question Usability Evaluation Toolkit*.
- Muhibuddin, A., Ishaqi, A., Triyana, I. W., Muhibuddin, A., Ishaqi, A., & Triyana, I. W. (2024). *Pengaruh Penerapan Asesmen Formatif dengan Metode Peer Assessment Terhadap Hasil Belajar Siswa pada Pembelajaran Matematika The Effect of Implementing Formative Assessment with the Peer Assessment Method on Student Learning Outcomes in Mathematics Learning*. 33(2), 547–556.
- Natsir, I., Mayasari, D., & Munfarikhatin, A. (2022). *Rasch Model : Quality of Test Instruments to Measure Students ' Ability in Algebra*. 01033, 6–10.
- Nisfatulsanah, A., Maret, U. S., Sugiharto, B., & Maret, U. S. (2024). *Comparing Classical Test Theory And Item Response*. 13(3), 70–86.
- Pastore, S. (2023). *Teacher assessment literacy : a systematic review*.
- Pratama, K. R., Mahmudah, I., Raya, I. P., & Tengah, K. (2024). Asesmen Formatif dalam Pembelajaran Matematika Memahami Perkembangan Siswa Secara Berkelanjutan Al-Ikram : Jurnal Pendidikan Guru Madrasah Ibtidaiyah 2 (1) (2024) : 1-8. *Al-Ikram Jurnal Pendidikan Guru Madrasah Ibtidaiyah*, 2(1), 15–23.
- Rohmatdi, A., Martono, M., Zafrullah, Z., & Safitri, R. (2024). *Optimizing Question Quality in Junior High School Exams : Classical Test Theory Evaluation with ITEMAN 4 . 3. 16*, 4363–4375. <https://doi.org/10.35445/alishlah.v16i4.5671>
- Risdianto, E., Wachidi, W., Riyanto, R., Alexon, A., Fathurrochman, I., & Kusen, K. (2021). Blended Learning Model Based on Massive Open Online Courses (MOOCs) Assisted by Augmented Reality (BMA) Model as the Electronic Learning Media in the Pandemic Covid-19. *Al-Ishlah Jurnal Pendidikan*, 13(1), 228–241. <https://doi.org/10.35445/alishlah.v13i1.470>
- Setyawati, R. D., Prasad, B., & Aini, S. N. (2023). Construct the Validity of STEM and Project-based Critical Thinking Skills Test Instruments Using the Rasch Model. *Phenomenon Jurnal Pendidikan Mipa*, 13(1), 96–110. <https://doi.org/10.21580/phen.2023.13.1.16246>
- Sumintono, B. (2018). Rasch Model Measurements as Tools in Assesment for Learning. <https://doi.org/10.2991/icei-17.2018.11>
- Susongko, P., Kusuma, M., Widiatmo, H., Tegal, U. P., & States, U. (2019). *Using Rasch Model to Detect Differential Person Functioning and Cheating Behavior in Natural Sciences Learning Achievement Test*. 5(2), 94–111.

<https://doi.org/10.30870/jppi.v5i2.5945>

- Töllner, F., & Kuhl, P. (2025). *Formative Assessment in Inclusive Mathematics Education in Secondary Schools : A Systematic Review*. 1–20.
- Uba, A. R. (2024). *Assessing Item Reliability , Differential Item Functioning , and Wright Map Analysis of the GSP122 Test at a Public University in Nigeria*. 2(2), 107–120.
- Ulwiyah, S. (2023). *Rasch Model Analysis On Mathematics Test Instruments : Biblioshiny (1983-2023)*. 7(2), 1–13.
- Widodo, W., & Chotimah, C. (2023). Adaptasi dan Analisis Psikometri Skala Kompetensi Multikultural Calon Guru Menggunakan Pemodelan Rasch. *Jurnal Pendidikan Dan Kebudayaan*, 8(2), 153–172. <https://doi.org/10.24832/jpnk.v8i2.4228>
- Wijaya, T. T. (2021). An Analysis of the Online Final Examination Items for Ninth-graders in the Mathematics Course Using the Rasch Measurement Model. <https://doi.org/10.20944/preprints202109.0510.v1>
- Yan, Z., & Pastore, S. (2022). Assessing Teachers' Strategies in Formative Assessment: The Teacher Formative Assessment Practice Scale. *Journal of Psychoeducational Assessment*, 40(5), 592–604. <https://doi.org/10.1177/07342829221075121>
- Yin, S., Chen, F., & Chang, H. (2022). *Assessment as Learning : How Does Peer Assessment Function in Students ' Learning ?* 13(June), 1–14. <https://doi.org/10.3389/fpsyg.2022.912568>