



Karakteristik Tingkat Kesulitan Butir Soal ASAS Matematika Kelas 3 SD/MI Berdasarkan Level Kognitif: Analisis Model Rasch

Nilna Masrurotus Sa'idah^{1*}, Sunismi²

^{1,2} Magister Pendidikan Matematika, Pascasarjana, Universitas Islam Malang

22502072003@unisma.com

Abstract

This study aims to describe the psychometric characteristics of the End-of-Semester Summative Assessment (ASAS) Mathematics items for Grade 3 elementary school based on three cognitive levels-knowledge and comprehension (L1), application (L2), and reasoning (L3) and to examine whether empirical item difficulty differs across these levels. Using a descriptive quantitative approach, responses from 81 third-grade students on 30 dichotomously scored items were analyzed with the Rasch model using Ministeps software. Results showed excellent item reliability (0.94; separation index 3.93) and adequate person reliability (0.78; separation index 1.86). Only one item (item 8) was identified as misfit, likely due to the dichotomization of a partial-credit complex multiple-choice item. The Wright Map indicated that the instrument was generally easy relative to students' ability, with item 17-classified as L2 (Application)-unexpectedly the most difficult item in the entire instrument, exceeding even the average difficulty of L3 items. Descriptive comparison showed that mean item difficulty followed the assumed order ($L1 \approx L2 < L3$), yet within-level variation, especially in L2, was substantial, indicating that the a priori cognitive-level classification in the ASAS blueprint does not fully and consistently reflect students' actual empirical item difficulty. These findings underscore the importance of empirical verification through the Rasch model as part of the routine item-review process for elementary mathematics assessments.

Keywords: Cognitive Level; Item Difficulty; Item Fit; Mathematics Assessment; Rasch Model

Abstrak

Penelitian ini bertujuan untuk mendeskripsikan karakteristik psikometrik butir soal Asesmen Sumatif Akhir Semester (ASAS) Matematika kelas 3 SD/MI berdasarkan tiga level kognitif pengetahuan dan pemahaman (L1), aplikasi (L2), dan penalaran (L3), serta menguji apakah tingkat kesulitan empiris butir berbeda antarlevel tersebut. Dengan pendekatan kuantitatif deskriptif, respons 81 siswa kelas 3 terhadap 30 butir soal yang dikodekan secara dikotomi dianalisis menggunakan Model Rasch berbantuan perangkat lunak Ministeps. Hasil penelitian menunjukkan reliabilitas item yang istimewa (0,94; indeks separasi 3,93) dan reliabilitas person yang memadai (0,78; indeks separasi 1,86). Hanya satu butir (butir nomor 8) yang teridentifikasi misfit, diduga akibat proses dikotomisasi pada butir pilihan ganda kompleks berskor parsial. Wright Map menunjukkan bahwa instrumen secara umum relatif mudah bagi kemampuan siswa, dengan butir nomor 17 yang diklasifikasikan sebagai L2 (Aplikasi) justru menjadi butir tersulit di seluruh instrumen, melampaui rata-rata tingkat kesulitan butir L3. Perbandingan deskriptif menunjukkan bahwa rata-rata tingkat kesulitan butir sejalan dengan urutan yang diasumsikan ($L1 \approx L2 < L3$), namun variasi tingkat kesulitan di dalam level, khususnya pada L2, cukup besar, yang mengindikasikan bahwa klasifikasi level kognitif secara apriori dalam kisi-kisi ASAS belum

sepenuhnya dan secara konsisten mencerminkan tingkat kesulitan empiris butir yang sesungguhnya dialami siswa. Temuan ini menegaskan pentingnya verifikasi empiris melalui Model Rasch sebagai bagian rutin dari proses telaah butir pada asesmen matematika sekolah dasar.

Kata Kunci: Asesmen Matematika; Kesesuaian Butir; Level Kognitif; Model Rasch; Tingkat Kesulitan Butir.

1. PENDAHULUAN

Penyusunan soal asesmen matematika di sekolah dasar umumnya disertai penetapan level kognitif yang menyesuaikan dengan taksonomi bloom revisi, yang membagi proses kognitif ke dalam enam dimensi, yaitu mengingat (C1), memahami (C2), mengaplikasikan (C3), menganalisis (C4), mengevaluasi (C5), dan mencipta (C6). Namun dalam prakteknya, keenam dimensi tersebut kerap disederhanakan ke dalam tiga kelompok, yakni level pengetahuan dan pemahaman (C1-C2), level aplikasi (C3), serta level penalaran atau berpikir tingkat tinggi (C4-C6) (Simarmata et al., 2024; Subiyantoro et al., 2020). Pengelompokan ini berangkat dari asumsi bahwa level yang lebih tinggi merepresentasikan tuntutan kognitif yang lebih kompleks, dari C1 sebagai level terendah hingga C6 sebagai level tertinggi. Namun, hubungan antara level kognitif yang dirancang dan tingkat kesulitan soal yang benar-benar dialami murid tidak selalu identik, sehingga penetapan level kognitif sebaiknya tidak hanya dilakukan secara apriori saat menyusun kisi-kisi. Masalah ini menjadi semakin krusial pada jenjang kelas rendah seperti kelas 3 SD/MI, di mana dasar numerasi murid masih dalam tahap pembentukan. Perkembangan numerasi dini terbukti memprediksi capaian matematika pada jenjang primary menengah, sehingga hasil belajar matematika di kelas-kelas awal dapat digunakan sebagai indikator untuk menilai kekuatan dasar numerasi murid. Temuan sejenis yang dipaparkan oleh Dewi dkk., (2025) juga menunjukkan bahwa kemampuan numerasi yang rendah sejak dini berkaitan dengan kesiapan sekolah dasar yang lebih lemah dan kesulitan menghadapi tantangan matematika berikutnya. Dalam konteks ini, penilaian merupakan bagian fundamental dari pembelajaran karena memberikan informasi tentang capaian belajar murid dan dapat digunakan untuk menyesuaikan pembelajaran secara lebih tepat (Chan & Oon, 2023; Purnamatati & Madani, 2023; Qadar et al., 2025). Pada pembelajaran matematika sekolah dasar, asesmen formatif yang dilakukan secara teratur juga dikaitkan dengan peningkatan pemahaman konsep, penguatan hasil belajar, dan adaptasi pembelajaran yang lebih dinamis

Penilaian merupakan bagian penting dalam proses pembelajaran karena digunakan untuk memperoleh informasi tentang capaian belajar murid dan menjadi dasar evaluasi hasil belajar. Dalam lingkungan sekolah, tes sering kali terlihat dalam bentuk Asesmen Sumatif Akhir Semester (ASAS), di mana hasilnya menjadi dasar untuk mengambil keputusan mengenai pencapaian belajar murid di akhir periode pembelajaran. Oleh karena itu, kualitas soal dalam instrumen tersebut perlu dijamin melalui bukti validitas, reliabilitas, keadilan, serta analisis karakteristik butir agar keputusan yang dihasilkan

akurat dan dapat dipertanggungjawabkan (Brownlie & Burke, 2026; Moralia & Namoco, 2025; Waty et al., 2024).

Dalam praktik penyusunan soal, tingkat kognitif biasanya ditetapkan dengan merujuk pada Taksonomi Bloom revisi yang mencakup enam dimensi proses kognitif, yaitu mengingat, memahami, mengaplikasikan, menganalisis, mengevaluasi, dan mencipta (Sümen & Karakaş, 2022). Untuk keperluan operasional, menurut Sümen & Karakaş, (2022), Simarmata et al., (2024) dan Subiyantoro et al., (2020) level tersebut sering disederhanakan menjadi tiga kategori, yaitu Level 1 (pengetahuan dan pemahaman), Level 2 (aplikasi), dan Level 3 (penalaran), dengan asumsi bahwa urutan tersebut mencerminkan peningkatan kompleksitas kognitif dan, dalam banyak praktik, juga diasosiasikan dengan peningkatan tingkat kesulitan soal. Namun, kemampuan guru atau calon guru dalam menyusun soal yang benar-benar sesuai dengan level kognitif yang ditargetkan belum selalu konsisten, soal pada level rendah cenderung lebih mudah disusun, sedangkan soal pada level aplikasi dan penalaran lebih sering menimbulkan kesalahan atau berada di bawah level yang seharusnya. Selain itu, analisis terhadap kisi-kisi dan ujian matematika menunjukkan bahwa klasifikasi awal guru dapat melebihi tuntutan kognitif soal dan tidak selalu selaras dengan hasil penelaahan ulang maupun performa empiris butir (Bartolome & Navalta, 2025). Dengan demikian, penentuan level kognitif dalam kisi-kisi soal yang dilakukan secara apriori perlu dilengkapi dengan verifikasi empiris berdasarkan data hasil pengerjaan murid dan analisis butir, agar tingkat kesulitan soal yang dialami murid benar-benar sesuai dengan level kognitif yang dimaksudkan (Chan & Oon, 2023; Moralia & Namoco, 2025).

Verifikasi empiris terhadap tingkat kesulitan soal dapat dilakukan melalui Model Rasch, karena model ini menganalisis hubungan antara kemampuan murid dan tingkat kesulitan butir serta telah banyak digunakan dalam analisis soal matematika pada berbagai jenjang pendidikan (Avrinda et al., 2024; Rizbudiani et al., 2021). Menurut Rizbudiani et al., (2021) dalam penelitian asesmen matematika, Model Rasch digunakan untuk menilai kecocokan butir, reliabilitas, dan sebaran tingkat kesulitan melalui perangkat lunak seperti QUEST, Winsteps, ConQuest, dan Ministeps. Penerapan berbantuan QUEST pada tes matematika menunjukkan bahwa kecocokan butir dapat dievaluasi melalui parameter INFIT MNSQ dan OUTFIT t; pada ujian akhir semester SMP, misalnya, 19 dari 20 butir dinyatakan layak dengan rentang INFIT MNSQ 0.89-1.17 dan hanya satu butir yang tidak memenuhi kriteria OUTFIT t. Studi lain yang dilakukan oleh Atikah et al., (2022) juga melaporkan bahwa seluruh 40 butir ujian matematika sesuai dengan model Rasch, dengan INFIT MNSQ 0.94-1.11 dan OUTFIT $t \leq 2.00$, serta tingkat kesulitan yang tersebar dari kategori mudah sampai sukar. Pada jenjang sekolah dasar, analisis Rasch dengan Winsteps telah digunakan untuk menilai validitas, reliabilitas, dan tingkat kesulitan soal matematika, baik pada instrumen kesiapan matematika siswa SD maupun pada instrumen kognitif matematika kelas tinggi. Analisis serupa pada butir matematika kelas V SD juga menunjukkan bahwa

Rasch dapat mengidentifikasi butir yang fit dan tidak fit berdasarkan Outfit MNSQ, Outfit ZSTD, dan Pt-measure Corr, sehingga hasilnya berguna untuk revisi instrumen (Aqida & Khairunisa, 2023).

Di sisi lain, penelitian pada kelas rendah sekolah dasar masih lebih sering menggunakan pendekatan teori tes klasik untuk menilai validitas, reliabilitas, tingkat kesukaran, daya beda, dan efektivitas distraktor (Hartono et al., 2023). Bahkan pada pengembangan instrumen matematika kelas 3 SD, fokus yang dilaporkan masih berada pada validitas, reliabilitas, asumsi model, dan pengukuran kemampuan umum murid, bukan pada analisis ASAS matematika menurut level kognitif L1, L2, dan L3. Karena itu, meskipun penelitian sebelumnya telah banyak menggunakan Model Rasch untuk menilai kualitas butir soal matematika dan sebagian telah mengelompokkan butir ke dalam kategori kognitif seperti LOTS, MOTS, dan HOTS, bukti yang secara spesifik menerapkan Model Rasch pada instrumen ASAS Matematika kelas 3 SD/MI serta menguji perbedaan tingkat kesulitan logit antarlevel kognitif secara statistik inferensial masih belum tampak dalam literatur yang tersedia (Qirom & Nurlaelah, 2023). Kesenjangan ini penting karena kelas 3 SD/MI merupakan fase penguatan fondasi numerasi, sehingga kesesuaian antara level kognitif yang dirancang dan tingkat kesulitan empiris butir perlu diuji agar instrumen ASAS dapat mengukur pencapaian belajar murid secara lebih akurat.

Berdasarkan paparan di atas, penelitian tentang analisis karakteristik butir soal ASAS Matematika kelas 3 SD/MI berdasarkan level kognitif L1, L2, dan L3 menggunakan Model Rasch dirasa sangat penting untuk dilaksanakan, dengan tujuan untuk mendeskripsikan karakteristik psikometrik butir soal yang mencakup tingkat kesulitan, kesesuaian item, dan reliabilitas serta untuk menguji apakah ada perbedaan signifikan secara statistik dalam tingkat kesulitan butir antar level kognitif tersebut, sehingga dapat diketahui apakah klasifikasi level kognitif yang terdapat dalam kisi-kisi ASAS telah mencerminkan tingkat kesulitan soal yang sebenarnya dihadapi oleh murid.

2. METODE PELAKSANAAN

Penelitian ini menerapkan metode kuantitatif deskriptif, yaitu sebuah pendekatan yang bertujuan untuk menggambarkan kondisi objek penelitian secara akurat berdasarkan data yang diperoleh, bukan untuk membuktikan atau membantah hipotesis. Metode ini digunakan untuk mengevaluasi mutu instrumen Asesmen Sumatif Akhir Semester (ASAS) Matematika untuk kelas 3 SD/MI berdasarkan tingkat kognitif L1 (pengetahuan dan pemahaman), L2 (aplikasi), dan L3 (penalaran) dengan menggunakan Model Rasch. Lokasi penelitian dilakukan di MI Muhammadiyah Manarul Islam Sawojajar Malang. Subjek dari penelitian ini terdiri dari 81 murid yang mengikuti ASAS Matematika kelas 3 pada Semester Genap Tahun Pelajaran 2025/2026, di mana pemilihan sampel menggunakan metode total sampling, sehingga semua peserta asesmen dijadikan subjek penelitian.

Instrument yang dianalisis adalah soal ASAS Matematika yang disusun oleh guru matematika di sekolah yang diteliti, berlandaskan capaian pembelajaran Matematika Fase B Kurikulum Merdeka dengan panduan soal ASAS yang sudah disepakati pada sekolah tersebut. Instrumen yang dianalisis adalah instrumen soal untuk kelas 3 yang terdiri dari 30 butir soal. Distribusi butir soal berdasarkan capaian pembelajarannya dan level kognitifnya bisa dilihat pada Tabel 1 berikut.

Tabel 1. Distribusi butir soal berdasarkan CP dan Level Kognitif

Capaian Pembelajaran	L1	L2	L3	Total
3.1 Bangun Datar	9	4	3	16
3.2 Pengolahan Data	6	4	4	14
Total	15	8	7	30

Sejalan dengan distribusi soal pada Tabel 1, dari ke-30 butir soal tersebut, pada kisi-kisi menjadi tiga tingkat kognitif dengan rincian 15 butir untuk level pengetahuan dan pemahaman (L1), 8 butir untuk level aplikasi (L2), dan 7 butir untuk level penalaran (L3). Soal tersebut hadir dalam berbagai bentuk seperti pada Tabel 2 berikut.

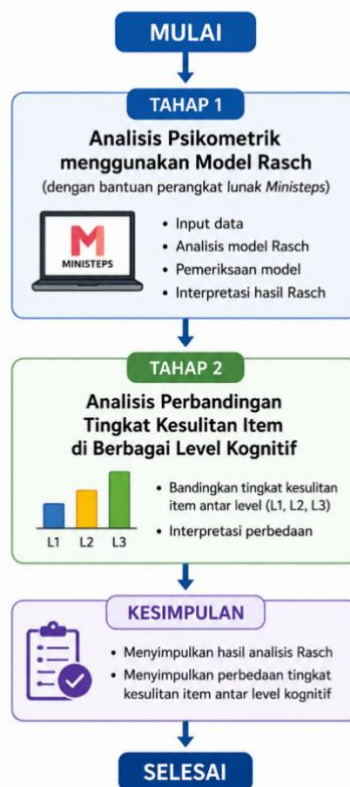
Tabel 2. Distribusi bentuk soal dan Level Kognitif

Bentuk Soal	L1	L2	L3	Total
Pilihan Ganda (PG)	5	0	0	5
Pilihan Ganda Kompleks (PGK)	3	1	0	4
Benar-Salah (BS)	1	1	1	3
Menjodohkan	2	1	0	3
Isian	4	5	1	10
Uraian	0	0	5	5
Total	15	8	7	30

Meskipun bentuk soal bervariasi, seluruh respons murid dikodekan secara dikotomi mengikuti kunci jawaban dan pedoman penskoran yang telah ditentukan, sehingga data yang dianalisis tetap konsisten dalam bentuk data dikotomi sesuai dengan yang dibutuhkan dalam Model Rasch.

Data penelitian berupa respons murid terhadap setiap butir soal yang dikodekan dengan skor 1 untuk jawaban yang benar dan skor 0 untuk jawaban yang salah. Data tersebut diperoleh melalui dokumentasi, yakni dengan meneliti dan mengumpulkan lembar jawaban serta hasil penskoran ASAS Matematika yang telah dilaksanakan. Dengan demikian, tidak ada perlakuan atau pengujian tambahan yang diberikan kepada murid selain tugas asesmen rutin tersebut.

Proses analisis data dilakukan dalam dua tahap yang disajikan dalam bagan alur berikut:



Gambar 1. Alur Analisis Data

Setelah butir soal dianalisis dan dinyatakan sesuai dengan Model Rasch jika nilai Infit/Outfit MNSQ berada dalam rentang 0,5–1,5, nilai ZSTD dalam rentang -2,0 sampai +2,0, dan nilai PTMEA Corr bernilai positif (Sumintono dan Widhiarso, 2015). Selain itu, distribusi kemampuan murid serta tingkat kesulitan butir disajikan dalam bentuk Wright Map untuk melihat kecocokan antara kemampuan murid dengan karakteristik soal. Selain itu, nilai logit untuk kesulitan setiap item yang dihasilkan dari Model Rasch akan dikelompokkan berdasarkan level kognitif L1 (pengetahuan dan pemahaman), L2 (aplikasi), dan L3 (penalaran). Kemudian, dilakukan analisis deskriptif dengan menggunakan rata-rata, simpangan baku, dan rentang logit untuk setiap level tersebut. Analisis deskriptif ini bertujuan untuk menilai apakah pengelompokan level kognitif yang ditetapkan dalam kisi-kisi ASAS sesuai dengan urutan kesulitan item yang sebenarnya dialami oleh para murid.

3. HASIL DAN PEMBAHASAN

3.1 Hasil

Estimasi Model Rasch dilakukan pada data respons dikotomi dari 81 siswa yang menjawab 30 butir soal ASAS Matematika. Tidak ditemukan skor ekstrem (0 atau skor sempurna) baik pada peserta maupun soal, sehingga semua 81 data siswa dan 30 butir soal dapat sepenuhnya digunakan dalam estimasi tanpa ada yang perlu dikeluarkan.

Tabel 3. Ringkasan Indeks Separasi dan Reliabilitas Model Rasch

	Indeks Separasi	Indeks Reliabilitas
Person (Siswa)	1,86	0,78
Item (Butir Soal)	3,93	0,94
KR-20 (Konsistensi Internal, Raw Score)	-	0,78

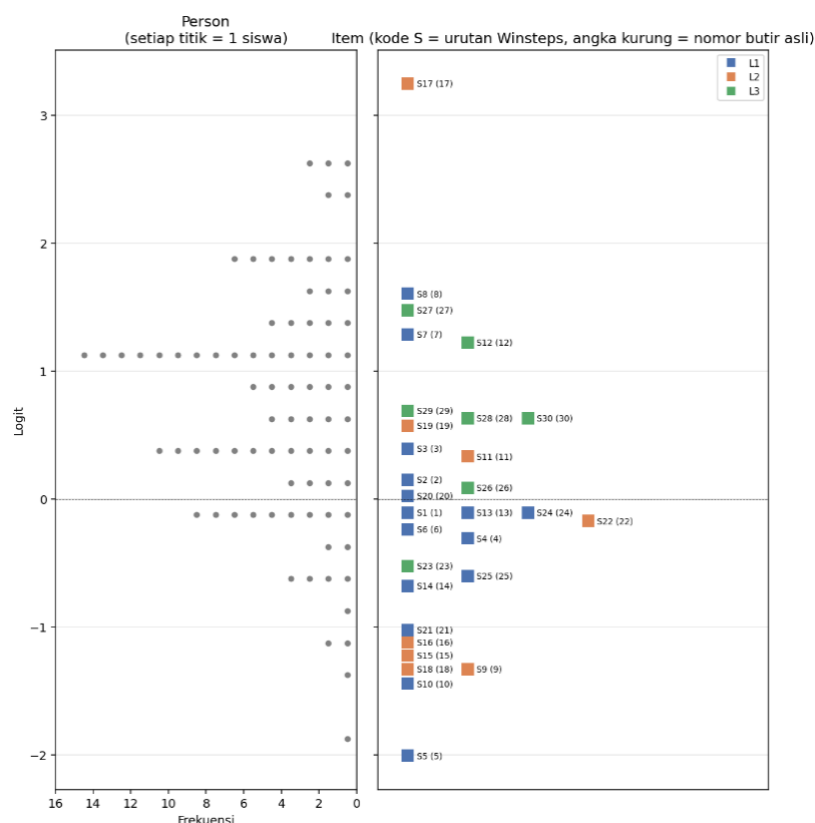
Tabel 3 menunjukkan bahwa indeks reliabilitas item mencapai 0,94 dengan separasi 3,93. Setelah butir soal dianalisis menggunakan Model Rasch, butir dinyatakan sesuai dengan model apabila memenuhi kriteria fit, yaitu nilai MNSQ berada pada rentang 0,5–1,5, nilai ZSTD berada pada rentang -2,0 sampai +2,0, dan nilai PT-Measure Corr bernilai positif, dengan kisaran yang lazim diterima 0,40–0,85 (Yulianto & Widodo, 2022). Selanjutnya, distribusi kemampuan murid dan tingkat kesulitan butir disajikan dalam bentuk Wright Map untuk memperlihatkan kesesuaian antara kemampuan murid dan karakteristik soal pada skala logit yang sama. Nilai logit kesulitan setiap butir kemudian dikelompokkan berdasarkan level kognitif L1, L2, dan L3, lalu dianalisis secara deskriptif menggunakan rata-rata, simpangan baku, dan rentang logit untuk menilai apakah pengelompokan level kognitif dalam kisi-kisi ASAS selaras dengan urutan kesulitan item yang secara empiris dialami murid (Azura et al., 2020).

Tabel 4. Estimasi Tingkat Kesulitan dan Kesesuaian (Fit) Butir Soal

No	Level	Logit	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	PTMEA Corr	Ket.
1	L1	-0.19	0.26	0.98	-0.06	1.03	0.22	0.41	Fit
2	L1	0.06	0.25	1.02	0.18	1.00	0.04	0.37	Fit
3	L1	0.31	0.24	1.07	0.50	1.05	0.37	0.33	Fit
4	L1	-0.39	0.26	1.02	0.20	1.13	0.85	0.35	Fit
5	L1	-2.10	0.41	1.03	0.25	0.78	-1.48	0.28	Fit
6	L1	-0.33	0.26	1.23	1.44	1.29	1.75	0.15	Fit
7	L1	1.20	0.25	0.88	-0.75	0.92	-0.50	0.48	Fit
8	L1	1.52	0.26	1.38	2.20	1.57	3.17	-0.03	Misfit
9	L2	-1.42	0.33	1.05	0.34	0.93	-0.43	0.29	Fit
10	L1	-1.53	0.34	0.75	-1.67	0.60	-2.98	0.56	Fit
11	L2	0.25	0.25	1.00	0.05	1.02	0.20	0.39	Fit
12	L3	1.14	0.25	0.78	-1.48	0.71	-2.01	0.59	Fit
13	L1	-0.19	0.26	1.02	0.20	0.93	-0.40	0.38	Fit
14	L1	-0.77	0.28	0.86	-0.90	0.73	-1.87	0.53	Fit
15	L2	-1.31	0.32	0.96	-0.23	0.90	-0.58	0.38	Fit
16	L2	-1.21	0.32	1.19	1.17	1.35	2.07	0.16	Fit
17	L2	3.33	0.41	1.07	0.49	1.09	0.62	0.17	Fit
18	L2	-1.42	0.33	0.91	-0.51	0.73	-1.85	0.43	Fit
19	L2	0.55	0.24	0.96	-0.23	0.98	-0.09	0.45	Fit
20	L1	0.37	0.24	0.98	-0.08	0.94	-0.33	0.43	Fit
21	L1	-0.54	0.27	0.90	-0.58	0.82	-1.18	0.49	Fit
22	L2	-0.26	0.26	0.92	-0.49	0.85	-0.99	0.48	Fit
23	L3	-0.54	0.27	0.96	-0.22	1.02	0.18	0.41	Fit
24	L1	-0.19	0.26	0.91	-0.55	0.83	-1.11	0.49	Fit
25	L1	-0.69	0.28	0.98	-0.11	1.04	0.33	0.39	Fit
26	L3	0.00	0.25	0.92	-0.48	0.85	-0.98	0.48	Fit
27	L3	1.45	0.26	1.03	0.21	1.18	1.13	0.32	Fit
28	L3	0.72	0.24	1.11	0.72	1.11	0.71	0.31	Fit
29	L3	0.72	0.24	1.21	1.31	1.32	1.91	0.19	Fit

No	Level	Logit	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	PTMEA Corr	Ket.
30	L3	1.45	0.26	0.93	-0.43	0.84	-1.04	0.45	Fit

Kriteria kesesuaian butir mengacu pada Yulianto & Widodo (2022), yaitu Infit/Outfit MNSQ pada rentang 0,5-1,5, ZSTD pada rentang -2,0 hingga +2,0, dan PTMEA Corr bernilai positif; suatu butir dinyatakan misfit apabila gagal memenuhi minimal dua dari kelima kriteria tersebut secara bersamaan. Berdasarkan kriteria tersebut, dari 30 butir soal hanya satu butir yang dinyatakan misfit terhadap Model Rasch, yaitu butir nomor 8 (Infit MNSQ = 1,38; Infit ZSTD = 2,20; Outfit MNSQ = 1,57; Outfit ZSTD = 3,17; PTMEA Corr = -0,03). Butir ini merupakan satu-satunya butir berbentuk pilihan ganda kompleks (PGK) berskor parsial (0-3) pada level L1. Setelah didikotomikan mengikuti pedoman penskoran, justru menunjukkan pola respons yang menyimpang dari model dan korelasi negatif dengan kemampuan siswa secara keseluruhan-mengindikasikan kemungkinan ambiguitas pada kunci jawaban parsial atau butir yang bersifat multitafsir bagi siswa kelas 3. Dua puluh sembilan butir lainnya, termasuk butir nomor 10, 12, dan 16 yang menunjukkan nilai ZSTD sedikit di luar rentang -2,0 pada salah satu indeks outfit/infit saja (overfit, MNSQ < 1), tetap dinyatakan fit terhadap Model Rasch karena hanya melanggar satu dari lima kriteria dan tidak menunjukkan ancaman berarti terhadap validitas pengukuran.



Gambar 2. Wright Map Sebaran Kemampuan Siswa dan Tingkat Kesulitan Butir Soal

Wright Map pada Gambar 2 menampilkan distribusi kemampuan siswa (sisi kiri) berdampingan dengan tingkat kesulitan butir (sisi kanan) pada skala logit yang sama. Rata-rata kemampuan siswa berada pada logit +0,62 (SD = 0,96), sedikit lebih tinggi dari rata-rata tingkat kesulitan butir yang dikalibrasi pada titik 0 logit, dengan 60 dari 81 siswa (74,1%) memiliki kemampuan di atas rata-rata kesulitan seluruh butir. Pola ini menunjukkan bahwa secara umum instrumen ASAS relatif mudah bagi mayoritas siswa kelas 3 yang menjadi subjek penelitian, meskipun sebaran butir masih cukup memadai untuk membedakan kemampuan siswa pada rentang kemampuan menengah hingga tinggi. Satu hal yang menonjol dari Wright Map adalah posisi butir nomor 17, yang terletak jauh di atas siswa dengan kemampuan tertinggi sekalipun (logit +3,33, sementara kemampuan siswa tertinggi hanya mencapai +2,68), dengan hanya 7 dari 81 siswa (8,6%) yang menjawab benar. Kesenjangan tersebut mengindikasikan bahwa butir nomor 17 kemungkinan terlalu sulit untuk mengukur kemampuan siswa kelas 3 secara proporsional, atau berpotensi mengandung permasalahan pada konstruksi maupun kunci jawabannya, mengingat tidak ada satu pun siswa yang kemampuannya cukup untuk menjawabnya dengan peluang di atas 50%.

Tabel 5. Tingkat Kesulitan Butir (Logit) Berdasarkan Level Kognitif

Level Kognitif	n Butir	Mean Logit	SD	Rentang (Min – Maks)
L1 (Pengetahuan dan Pemahaman)	15	-0,23	0,91	-2,10 s.d. 1,52
L2 (Aplikasi)	8	-0,19	1,62	-1,42 s.d. 3,33
L3 (Penalaran)	7	0,71	0,75	-0,54 s.d. 1,45

Tabel 5 menunjukkan bahwa urutan rata-rata logit kesulitan butir antarlevel sudah sesuai arah yang diasumsikan secara apriori, yaitu L1 dan L2 relatif setara (masing-masing -0,23 dan -0,19 logit) dan lebih mudah dibandingkan L3 (0,71 logit). Meskipun demikian, sebaran logit pada L2 jauh lebih lebar (SD = 1,62; rentang -1,42 s.d. 3,33) dibandingkan L1 (SD = 0,91; rentang -2,10 s.d. 1,52) dan L3 (SD = 0,75; rentang -0,54 s.d. 1,45), yang terutama disebabkan oleh satu butir dengan logit ekstrem, yaitu butir nomor 17. Variasi tingkat kesulitan di dalam masing-masing level yang cukup besar-khususnya pada L2-mengindikasikan bahwa perbedaan rata-rata antarlevel tersebut perlu dimaknai secara hati-hati, mengingat jumlah butir pada tiap level yang relatif sedikit (7-15 butir) turut membatasi kekuatan generalisasi dari perbandingan rata-rata semata.

3.2 Pembahasan

Secara keseluruhan, hasil dari analisis Model Rasch menunjukkan bahwa instrumen ASAS Matematik untuk kelas 3 SD/MI yang diperiksa memiliki kualitas psikometrik yang baik, dengan reliabilitas item yang sangat baik (0,94) dan reliabilitas person yang cukup (0,78). Temuan ini sejalan dengan yang dilaporkan oleh Sheptian et al., (2024) dan Yudha (2023) bahwa reliabilitas item dan person dalam kategori baik hingga luar biasa ketika soal matematika sekolah dianalisis menggunakan Model Rasch dengan bantuan Winsteps, yang membuktikan bahwa pendekatan Rasch secara konsisten dapat

memberikan estimasi psikometrik yang stabil pada berbagai tingkat dan jenis instrumen matematika sekolah (Rizbudiani et al., 2021).

Dalam hal kecocokan soal, hanya satu dari 30 soal, yaitu nomor 8, yang dianggap tidak cocok dengan Model Rasch. Temuan ini menguatkan hasil penelitian Effatpanah et al., (2024) yang juga menemukan adanya soal yang tidak cocok pada instrumen matematika di sekolah menengah berdasarkan nilai Infit dan Outfit, meskipun perangkat lunak yang digunakan berbeda (QUEST dalam penelitian tersebut, dibandingkan dengan estimasi Rasch dalam studi ini). Karakteristik khusus dari soal nomor 8, sebagai satu-satunya soal tipe pilihan ganda kompleks (PGK) dengan skor parsial yang kemudian didikotomikan sesuai dengan panduan penilaian, menunjukkan bahwa proses dikotomisasi pada soal berskor parsial berisiko menyebabkan hilangnya informasi dari respons siswa. Ini mengakibatkan pola jawaban yang sebenarnya bertahap (credit parsial) terpaksa diubah menjadi pola benar-salah yang tidak selalu sesuai dengan asumsi model Rasch yang dikotomi (May et al., 2023). Ini menjadi catatan penting bagi penyusun ASAS agar mempertimbangkan penggunaan Model Rasch Partial Credit bukannya model dikotomi murni jika instrumen di masa depan tetap menyertakan soal berskor parsial, serta melakukan tinjauan kualitatif ulang terhadap kunci jawaban dan panduan penilaian soal nomor 8.

Wright Map pada Gambar 2 memberikan gambaran tambahan yang tidak terlihat hanya dari angka reliabilitas atau statistik fit, yaitu kecocokan antara distribusi kemampuan siswa dan tingkat kesulitan soal. Rata-rata kemampuan siswa (+0,62 logit) yang lebih tinggi daripada rata-rata kesulitan soal (0 logit, berdasarkan konvensi kalibrasi) menunjukkan bahwa instrumen ASAS ini umumnya cukup mudah bagi sebagian besar siswa kelas 3 di sekolah tempat penelitian dilakukan, yang merupakan pola yang wajar untuk asesmen akhir semester yang dirancang agar banyak siswa dapat mencapai kompetensi minimal. Namun, keberadaan soal nomor 17 yang tingkat logitnya jauh melebihi kemampuan siswa tertinggi sekalipun adalah temuan yang perlu diperhatikan lebih lanjut. Soal tersebut diklasifikasikan sebagai L2 (Aplikasi) dalam kisi-kisi ASAS, tetapi secara empiris menjadi soal yang paling sulit di antara instrumen tersebut, jauh melebihi rata-rata logit tujuh soal L3 (Penalaran) sekalipun. Ketidaksesuaian ini menjadi bukti yang paling nyata dari perbedaan antara tingkat kognitif yang ditentukan sebelumnya oleh guru dan tingkat kesulitan yang dialami siswa (Rahmatillah & Zainuddin, 2025; Rizbudiani et al., 2021). Dengan demikian, hasil ini mendukung perlunya verifikasi empiris terhadap klasifikasi level kognitif soal, khususnya pada instrumen ASAS matematika kelas rendah, agar keputusan penilaian lebih akurat dan sesuai dengan kemampuan nyata murid.

Karakteristik soal di berbagai tingkat kognitif terlihat dari jenis soal yang digunakan (Tabel 2). Soal-soal di L1 lebih banyak menggunakan jenis soal tertutup dengan satu jawaban, seperti pilihan ganda (5 soal) dan pilihan ganda kompleks (3 soal), yang

biasanya mengukur proses ingat dan paham secara langsung. Di L2, lebih banyak soal berbentuk isian (5 dari 8 soal), yang meminta siswa untuk menerapkan suatu prosedur atau konsep dalam Situasi tertentu tanpa perlu menjelaskan alasannya secara jelas. Sementara di L3, tidak ada soal berupa pilihan ganda; lima soalnya adalah soal uraian yang mengharuskan siswa menuliskan proses berpikirnya secara jelas, sesuai dengan tuntutan proses berpikir menganalisis, mengevaluasi, dan mencipta (C4-C6) pada tingkat penalaran. Pola ini menunjukkan bahwa kerumitan bentuk soal cenderung meningkat sejalan dengan tingkat kognitif, yang juga berkontribusi pada perbedaan tingkat kesulitan antar tingkat, di luar materi soal itu sendiri (Yulianto & Widodo, 2022).

Namun, data tentang tingkat kesulitan pada setiap soal (Tabel 4) menunjukkan bahwa label tingkat kognitif tidak selalu konsisten dalam memprediksi tingkat kesulitan nyata dari tiap soal. Di L1, misalnya, beberapa soal dianggap sangat mudah seperti soal nomor 5 (-2,10 logit) dan soal nomor 10 (-1,53 logit), sementara beberapa lainnya tergolong sulit, seperti soal nomor 7 (1,20 logit) dan soal nomor 8 (1,52 logit, misfit), yang lebih sulit dibandingkan beberapa soal di L3 seperti soal nomor 23 (-0,54 logit) dan soal nomor 26 (0,00 logit). Pola yang sama terjadi pada L2, yang memiliki baik soal-soal termudah dalam instrumen (soal nomor 9, 15, 16, dan 18, semuanya di bawah -1,20 logit) maupun soal tersulit dalam instrumen (soal nomor 17, 3,33 logit), sehingga L2 menjadi tingkat dengan variasi kesulitan soal yang paling beragam. Berbeda dengan L1 dan L2, ketujuh soal di L3 menunjukkan penyebaran yang lebih seragam dan konsisten berada di kisaran kesulitan sedang hingga sulit (-0,54 hingga 1,45 logit), membuat L3 menjadi tingkat yang paling konsisten mencerminkan asumsi bahwa penalaran lebih sulit daripada pengetahuan dan penerapan. Dengan demikian, dapat disimpulkan bahwa perbedaan tingkat kesulitan antar tingkat kognitif di instrumen ini lebih banyak dipengaruhi oleh variasi kualitas pembuatan tiap soal daripada hanya oleh label tingkat kognitifnya, khususnya di L1 dan L2, sementara label tingkat kognitif di L3 lebih dapat diandalkan untuk memprediksi tingkat kesulitan nyata dari soal (Bartolome & Navalta, 2025; Yulianto & Widodo, 2022).

Pada perbandingan kelompok, nilai rata-rata kesulitan L1 (-0,23) dan L2 (-0,19) terlihat hampir sama dan keduanya lebih mudah daripada L3 (0,71), sesuai dengan urutan yang sudah diharapkan dalam kisi-kisi ASAS. Namun, penyebaran nilai di setiap level-khususnya di L2 yang memiliki simpangan baku ($SD = 1,62$) jauh lebih tinggi dibandingkan L1 ($SD = 0,91$) dan L3 ($SD = 0,75$) akibat butir nomor 17 yang merupakan nilai ekstrim-menunjukkan bahwa perbedaan rata-rata antara level tersebut tidak selalu menunjukkan pemisahan yang jelas dan konsisten di antara ketiga level kognitif. Penemuan ini melengkapi penelitian sebelumnya seperti Yulianto & Widodo (2022) yang mendeskripsikan klasifikasi butir Penilaian Akhir Semester Matematika ke dalam kategori LOTS, MOTS, dan HOTS, dengan menambahkan catatan bahwa perbandingan rata-rata logit antarlevel saja belum cukup untuk menyimpulkan adanya pemisahan level kesulitan yang jelas, mengingat ada variasi yang besar di dalam level tersebut dan

jumlah butir pada setiap level yang cukup sedikit (7 hingga 15 butir). Ketidakjelasan pemisahan ini bisa dilihat dari dua sisi. Pertama, secara substansial, hasil ini menunjukkan bahwa pembagian level kognitif L1, L2, dan L3 yang ditetapkan oleh guru dalam kisi-kisi ASAS belum sepenuhnya mencerminkan tingkat kesulitan yang sebenarnya dihadapi siswa secara bertahap dan konsisten, seperti yang diidentifikasi pada bagian Pendahuluan. Kedua, secara metodologis, jumlah butir yang sedikit di setiap level membatasi kemampuan untuk membuat generalisasi dari pola rata-rata yang terlihat, sehingga hasil ini sebaiknya dianggap sebagai indikasi awal yang perlu dikonfirmasi lebih lanjut dengan instrumen yang memiliki jumlah butir di setiap level yang lebih banyak.

Secara keseluruhan, temuan tersebut menekankan pentingnya verifikasi empiris untuk klasifikasi level kognitif yang ditetapkan sebelumnya dalam penyusunan kisi-kisi ASAS, seperti yang disebutkan di Pendahuluan. Meskipun secara rata-rata hanya ada sedikit perbedaan antarlevel yang menunjukkan pemisahan yang jelas dan konsisten, temuan pada level butir individu-baik berupa butir yang tidak sesuai (nomor 8) maupun butir yang kesulitannya jauh berbeda dari label kognitifnya (butir nomor 17)-menunjukkan bahwa analisis Model Rasch dapat mengungkap informasi diagnostic yang tidak bisa diperoleh hanya dari perbandingan rata-rata logit antarlevel. Ini memperkuat argumen bahwa verifikasi psikometrik yang berbasis Model Rasch perlu menjadi bagian rutin dari proses penyusunan dan telaah butir ASAS, terutama pada tingkat kelas rendah seperti kelas 3 SD/MI yang menjadi dasar pemahaman numerasi siswa, sehingga perbedaan antara tujuan pengajaran penyusun soal dan pengalaman nyata siswa dalam mengerjakan soal dapat diminimalkan.

4. SIMPULAN

Berdasarkan hasil analisis Model Rasch terhadap 30 butir soal ASAS Matematika kelas 3 SD/MI yang dikerjakan oleh 81 siswa, dapat disimpulkan bahwa instrumen tersebut memiliki kualitas psikometrik yang baik, ditunjukkan oleh reliabilitas item yang istimewa (0,94; separasi 3,93) dan reliabilitas person yang memadai (0,78; separasi 1,86), serta hanya satu dari 30 butir (butir nomor 8) yang dinyatakan misfit terhadap model, diduga akibat proses dikotomisasi pada butir berskor parsial. Secara keseluruhan, instrumen relatif mudah bagi mayoritas siswa, dengan rata-rata kemampuan siswa (+0,62 logit) berada di atas rata-rata tingkat kesulitan butir. Perbandingan tingkat kesulitan butir antarlevel kognitif menunjukkan bahwa urutan rata-rata logit L1, L2, dan L3 secara umum sejalan dengan asumsi apriori (L1 dan L2 relatif setara dan lebih mudah dibandingkan L3), namun variasi tingkat kesulitan di dalam masing-masing level khususnya L2, yang memiliki sebaran logit paling lebar akibat butir nomor 17 menunjukkan bahwa klasifikasi level kognitif dalam kisi-kisi ASAS belum sepenuhnya konsisten mencerminkan tingkat kesulitan empiris yang sesungguhnya dialami siswa. Temuan pada butir nomor 8 (misfit) dan butir nomor 17 (kesulitan empiris jauh melampaui label kognitifnya) menegaskan bahwa verifikasi empiris berbasis Model

Rasch mampu mengungkapkan informasi diagnostik penting yang tidak dapat diperoleh hanya dari penetapan level kognitif secara apriori.

5. UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Kepala MI Muhammadiyah Manarul Islam Sawojajar Malang beserta guru mata pelajaran matematika kelas 3 atas izin dan bantuan yang diberikan selama proses pengumpulan data penelitian ini. Ucapan terima kasih juga disampaikan kepada seluruh siswa kelas 3 yang telah mengikuti Asesmen Sumatif Akhir Semester (ASAS) Matematika sebagai sumber data penelitian, serta kepada dosen pembimbing atas arahan dan masukan yang berharga dalam penyusunan artikel ini.

6. REKOMENDASI

Berdasarkan temuan penelitian, disarankan agar penyusun ASAS Matematika di sekolah dasar tidak hanya mengandalkan penetapan level kognitif secara apriori pada saat penyusunan kisi-kisi, tetapi juga melakukan verifikasi empiris terhadap tingkat kesulitan butir melalui analisis Model Rasch, khususnya untuk butir pada level aplikasi (L2) yang pada penelitian ini menunjukkan variasi tingkat kesulitan paling besar. Butir nomor 8 yang teridentifikasi misfit perlu ditelaah ulang secara kualitatif, terutama pada kunci jawaban dan pedoman penskoran parsialnya, dan penyusun instrumen disarankan mempertimbangkan penggunaan Model Rasch Partial Credit apabila ke depan tetap menyertakan butir berskor parsial agar informasi respons siswa tidak hilang akibat proses dikotomisasi. Butir nomor 17 yang secara empiris jauh lebih sulit dibandingkan kemampuan seluruh siswa perlu ditinjau ulang, baik dari sisi konstruksi soal maupun kesesuaiannya dengan label level kognitif Aplikasi (L2) yang ditetapkan. Bagi peneliti selanjutnya, disarankan untuk melakukan penelitian serupa dengan jumlah butir yang lebih banyak pada setiap level kognitif agar perbandingan tingkat kesulitan antarlevel dapat digeneralisasi dengan lebih kuat, serta mempertimbangkan penggunaan uji statistik inferensial (seperti ANOVA satu jalur pada nilai logit) untuk menguji signifikansi perbedaan tingkat kesulitan antarlevel kognitif secara lebih ketat.

7. REFERENSI

- Aqida, D. S., & Khairunisa, D. (2023). *The Use Of Rasch Model In Analyzing Grade V Lementary School Mathematics Items*. 7(1), 57–67.
- Atikah, Sudiyatno, Rahim, A., & Marlina. (2022). Assessing the item of final assessment mathematics test of junior high school using Rasch model. *Jurnal Elemen*, 8(1), 117–130.
- Avrinda, C. A., EN, A. S., & Ghani, F. A. (2024). Analyzing mathematics formative test questions for higher-order thinking skills utilizing the Rasch model. *Journal of Teaching and Learning Mathematics*, 130–140.

- Azura, A., Samsudin, A., & Utari, S. (2020). Analisis Peta Wright Keterampilan Berpikir Level Lots Dan Hots Siswa Kelas Xi Pada Materi Miopi. *WaPFI (Wahana Pendidikan Fisika)*, 5(1), 76–83.
- Bartolome, D. J. P., & Navalta, J. D. (2025). *Misalignment in Cognitive Levels : A Taxonomic Analysis of Tables of Specification and Mathematics Examinations*. 2(2), 194–207.
- Brownlie, N., & Burke, K. (2026). *Quality indicators of effective summative assessment*. July. <https://doi.org/10.1108/QAE-04-2023-0062>
- Chan, W. K., & Oon, E. P. (2023). *Classroom assessment that tailor instruction and direct learning : A validation study*. 10(2), 376–394.
- Dewi, N. K., Pudyaningtyas, A. R., Sholeha, V., Rahmawati, A., Palupi, W., & Syamsuddin, M. M. (2025). Early Childhood Numeracy Skill Assessment : A Preparation Study For Elementary School Transition. *KIDDO : JURNAL PENDIDIKAN ISLAM ANAK USIA DINI*, 6(27), 27–42.
- Effatpanah, F., Baghaei, P., Tabatabaee-Yazdi, M., & Babaii, E. (2024). A new scoring method for item response theory analysis of C-Tests. *Language Testing*, 42, 167–192. <https://doi.org/10.1177/02655322241265350>
- Hartono, W., Hadi, S., Rosnawati, R., & Retnawati, H. (2023). Exploration of Student ' Cognitive Mathematics Ability Diagnostic Instruments : Validity , Reliability , and Item Characteristics. *Pege Journal of Education and Instruction*, 13(3), 386–394. <https://doi.org/10.47750/pegegog.1>
- May, T., Koskey, K., Bostic, J., Stone, G., Kruse, L., & Matney, G. (2023). Examining how using dichotomous and partial credit scoring models influence sixth-grade mathematical problem-solving assessment outcomes. *School Science and Mathematics*. <https://doi.org/10.1111/ssm.12570>
- Moralia, M. J. A., & Namoco, S. O. (2025). Validity and Reliability of Summative Test Assessment for Bachelor of Technology and Livelihood Education Students. *Technium Social Sciences Journal*, 74(1), 96–112. <https://doi.org/10.47577/tssj.v74i1.13108>
- Purnamatati, A. M. M., & Madani, F. (2023). Analisis Assesmen Autentik Pembelajaran Matematika Sekolah Dasar. 6(2), 778–788. <https://doi.org/10.31949/jee.v6i2.5659>
- Qadar, R., Efwinda, S., Syam, M., Putri, R., Haerani, R., Dinurrohman, S., Aryaputra, R., Devina, S., Putri, W., Mulawarman, U., & Mulawarman, U. (2025). *Assessment of Competence of Pre-service Mathematics and Science Teachers in Classifying Cognitive Processes and Knowledge Dimensions through the Revised Bloom's Taxonomy*. 5(2), 248–256.
- Qirom, M. S., & Nurlaelah, E. (2023). Rasch Model : Analyzing the Items Quality of Mathematics Higher- Order Thinking Skill Instrument. *Jurnal Pendidikan Dan Pembelajaran*, 30(1), 8–16.
- Rahmatillah, N., & Zainuddin, A. (2025). Evaluation of HOTS Test on Renewable Energy Materials through Rasch Model Analysis. *Jurnal Pijar MIPA*, 20(5), 933–939.
- Rizbudiani, A. D., Jaedun, A., & Rahim, A. (2021). Rasch model item response theory (IRT) to analyze the quality of mathematics final semester exam test on system of linear equations in two variables (SLETV). *Al-Jabar: Jurnal Pendidikan Matematika*, 12(2), 399–412.
- Sheptian, R., Arthur, R., & Wardhani, R. (2024). Psychometric Analysis Of Mathematics Readiness Assessment Tool For Elementary School Students: A Rasch Model Approach. *LEARNING : Jurnal Inovasi Penelitian Pendidikan Dan Pembelajaran*, 4(4), 1213–1220.
- Simarmata, G., Simanungkalit, R. H., & Sinabang, F. S. (2024). *Analisis Tingkat Kognitif Soal*

- pada Buku Kemdikbudristek 2022 Materi Teorema Pythagoras. 8(3), 517–526.*
- Subiyantoro, E., Ashari, A., & Suprpto. (2020). Cognitive Classification Based on Revised Bloom's Taxonomy Using Learning Vector Quantization. *2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, 349–353. <https://doi.org/10.1109/CENIM51130.2020.9297879>
- Sümen, Ö. Ö., & Karakaş, H. (2022). Analysis of the Cognitive Levels of the Pre-service Primary Teachers' Questions on Fractions According to the Revised Bloom Taxonomy. *Cumhuriyet International Journal of Education, 11(3)*, 512–521.
- Waty, E. R. K., Nengsih, Y. K., & Handrianto, C. (2024). The quality of teacher-made summative tests for Islamic education subject teachers in Palembang, Indonesia. *Cakrawala Pendidikan, 43(1)*, 192–203.
- Yudha, R. P. (2023). Higher Order Thinking Skills (HOTS) Test Instrument: Validity and Reliability Analysis With The Rasch Model. *Mathematics Education Learning And Teaching, 12(1)*, 21–38.
- Yulianto, A., & Widodo, A. (2022). Disclosure of Difficulty Distribution of HOTS-Based Test Questions through Rasch Modeling. *Indonesian Journal of Primary Education*. <https://consensus.app/papers/indonesian-journal-of-primary-education-disclosure-of-yulianto-widodo/ec4f322b466657fa983086abfebf5949/>