

Evaluasi Pembelajaran: Menafsir Ulang 'Rendah' dalam TKA Matematika - Sebuah Kritik Metodologis terhadap Pengukuran Kompetensi Abad 21

Eli Agustinah^{a,1}, Sukiyanto^{b,2,*}

^a Universitas Sarjanawiyata Tamansiswa Jogjakarta;

^{*b} Universitas Sarjanawiyata Tamansiswa Jogjakarta;

¹ aishaummkhurayya@gmail.com; ² sukiyanto.math@ustjogja.ac.id;

*Correspondent Author; sukiyanto.math@ustjogja.ac.id

ARTICLE INFO

Article history

Received:

30-05-2026

Revised:

14-06-2026

Accepted:

01-08-2026

Keywords

Item Response Theory (IRT), Kompetensi Abad 21, Kritik Metodologis, TKA Matematika, Assessment Artifacts, Differential Item Functioning (DIF).

ABSTRACT

Penafsiran tunggal terhadap rerata skor Tes Kemampuan Akademik (TKA) Matematika yang berada pada kategori "rendah" sering kali secara prematur disimpulkan sebagai krisis kompetensi siswa, tanpa mengevaluasi kualitas psikometrik instrumen yang digunakan. Penelitian ini bertujuan untuk melakukan kritik metodologis dan menafsirkan ulang label skor "rendah" melalui analisis psikometrik modern dan dekonstruksi keselarasan konstruk kompetensi abad ke-21 (4C). Penelitian kuantitatif deskriptif berdesain evaluatif-kritikal ini menganalisis respons sekunder 1.250 siswa SMP/MTs di Kabupaten Malang terhadap 40 butir soal TKA Matematika berbasis Higher-Order Thinking Skills (HOTS). Analisis data menggunakan Item Response Theory (IRT) model Rasch (1-PL) dan Two-Parameter Logistic (2-PL) untuk menguji item fit, tingkat kesukaran (b), daya beda (a), serta Differential Item Functioning (DIF), yang dilanjutkan dengan pemetaan konten-kritis terhadap indikator 4C. Hasil penelitian menunjukkan bahwa 30,0% butir soal dikategorikan misfit, 37,5% memiliki kesukaran sangat tinggi ($b > 2,0$), 22,5% berdaya beda buruk ($a < 0,5$), dan 27,5% terbukti memuat bias lokasi (DIF) yang merugikan siswa dari daerah rural. Selain itu, hanya 42,5% butir soal yang memiliki keselarasan konstruk secara valid dengan indikator 4C, di mana mayoritas soal yang dilabeli HOTS sejatinya hanya menguji kompleksitas kalkulasi (procedural complexity), bukan kedalaman penalaran kritis (Critical Thinking) atau fleksibilitas strategi (Creativity). Temuan ini membuktikan bahwa label skor "rendah" pada TKA Matematika bukan murni mencerminkan ketidakmampuan akademis siswa (true low competence), melainkan akibat dari artefak pengukuran (assessment artifacts) dan kecacatan desain instrumen. Penelitian ini merekomendasikan penghentian praktik menyalahkan kapasitas siswa (blaming the learner) dan mendesak digelarnya kalibrasi psikometrik serta standardisasi instrumen tes sebelum instrumen dijadikan dasar formulasi kebijakan pendidikan.

This is an open-access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Introduction

Pendidikan global saat ini dihadapkan pada disrupsi besar yang menuntut reorientasi pengukuran hasil belajar, dari sekadar hafalan prosedural menuju penguasaan kompetensi abad ke-21 yang mencakup penalaran tingkat tinggi (Higher-Order Thinking Skills / HOTS), pemecahan masalah kompleks, dan literasi numerasi. Namun, instrumen evaluasi pembelajaran khususnya Tes Kemampuan Akademik (TKA) Matematika sering kali masih ditafsirkan secara dangkal berbasis skor mentah tunggal. Ketika rerata skor peserta didik berada di kategori "rendah", kebijakan publik dan publikasi media secara prematur langsung mengklaim terjadinya krisis kompetensi matematis. Cara pandang yang reduksionistik ini memicu bias pengukuran, karena melupakan analisis konstruksi instrumen, validitas item, serta disparitas konteks sosial-ekologis tempat pembelajaran berlangsung. Reduksi klaim "rendah" tanpa kritik metodologis tidak hanya menyesatkan evaluasi kurikulum, tetapi juga berisiko menghasilkan kebijakan intervensi pedagogis yang salah sasaran.

Fenomena klaim "rendah" dalam asesmen matematika secara konsisten muncul di berbagai tingkatan. Pada skala internasional, laporan Programme for International Student Assessment (PISA) merilis bahwa sekitar 31% siswa di negara-negara anggota OECD belum mencapai Tingkat 2 (tingkat kemahiran minimum) dalam literasi matematika (OECD, 2023). Di tingkat nasional, kondisi ini semakin mengkhawatirkan di mana hasil PISA menunjukkan bahwa 82% siswa Indonesia berada di bawah tingkat kecakapan minimum matematika, memperlihatkan stagnasi kemampuan penalaran numerasi yang berkepanjangan (Kemendikbudristek, 2023). Sementara itu pada skala regional, khususnya di Kabupaten Malang, Jawa Timur, asesmen daerah menunjukkan bahwa hanya 38,4% siswa SMP/MTs yang mampu memenuhi standar ketuntasan minimum pada domain soal matematika berbasis penalaran dan HOTS (Rizky et al., 2023). Angka-angka ini kerap dilabeli secara dramatis sebagai kegagalan fatal peserta didik, padahal ada indikasi kuat bahwa rendahnya performa tersebut dipicu oleh ketidaksesuaian (mismatch) antara instrumen asesmen formal dengan paradigma penilaian kompetensi abad ke-21.

Urgensi penelitian ini terletak pada mendesaknya kebutuhan untuk mendekomposisi dan menafsirkan ulang label "rendah" dalam TKA Matematika sebelum dijadikan dasar pengambilan keputusan strategis. Mengaitkan skor rendah secara langsung dengan kegagalan intelektual siswa tanpa mengevaluasi kualitas psikometrik soal, daya beda, serta tingkat bias budaya instrumen dapat merusak ekosistem pembelajaran. Tanpa re-evaluasi kritis terhadap cara kita mengukur, kebijakan perbaikan kurikulum matematika akan terus terjebak dalam lingkaran setan *teaching to the test* (mengajar hanya demi lulus ujian), yang justru melumpuhkan daya kritis dan kreativitas siswa yang dibutuhkan di abad ke-21.

Meskipun kajian mengenai rendahnya prestasi matematika telah banyak dilakukan, terdapat gap penelitian (research gap) yang nyata dalam literatur evaluasi pendidikan saat ini. Sebagian besar studi terdahulu (seperti Aini & Suryo, 2021; Wardani et al., 2022) cenderung mengasumsikan instrumen TKA sebagai alat ukur yang absolut valid dan berfokus secara eksklusif pada faktor internal siswa (seperti kecemasan matematika atau gaya belajar) sebagai penyebab skor rendah. Sangat sedikit penelitian yang mengambil sudut pandang kebalikannya: meneliti instrumen asesmen itu sendiri melalui kacamata kritik metodologis untuk menguji apakah label "rendah" tersebut merupakan hasil dari cacat desain pengujian (assessment artifacts) atau kegagalan konstruksi instrumen dalam memotret indikator kompetensi abad ke-21.

Kebaruan (novelty) penelitian ini menghadirkan kerangka dekonstruksi metodologis baru yang memadukan teori evaluasi psikometrik modern (Item Response Theory) dengan dimensi kompetensi abad ke-21 (4C: Critical Thinking, Creativity, Collaboration, Communication). Berbeda dari penelitian konvensional yang memosisikan angka statistik

TKA sebagai vonis akhir, penelitian ini membedah struktur soal TKA Matematika untuk mengidentifikasi Differential Item Functioning (DIF) dan merumuskan ulang re-interpretasi kuantitati-kualitatif atas atribut "rendah". Pendekatan ini menghasilkan cara pandang evaluatif baru yang menggeser prasangka "siswa tidak mampu" menjadi evaluasi "instrumen yang tidak relevan".

Alasan utama pengambilan judul "Evaluasi Pembelajaran: Menafsir Ulang 'Rendah' dalam TKA Matematika - Sebuah Kritik Metodologis terhadap Pengukuran Kompetensi Abad 21" adalah untuk memberikan kontribusi korektif atas maraknya simplifikasi diagnostik pendidikan di Indonesia. Judul ini dipilih secara sengaja untuk memicu paradigma baru bagi para pendidik, pembuat kebijakan, dan pengembang tes agar tidak lagi menjadikan skor TKA sebagai "hakim tunggal" kompetensi matematika. Melalui judul ini, penelitian ini bertujuan meluruskan arah evaluasi pembelajaran agar lebih akurat, adil, dan benar-benar mencerminkan kapasitas penalaran matematika siswa di era modern.

Evaluasi pembelajaran matematika merupakan proses pengumpulan, pemrosesan, dan interpretasi bukti secara terstruktur mengenai perkembangan kapasitas kognitif serta afektif siswa. Dalam lingkup disiplin ilmu matematika, evaluasi tidak boleh dipersempit sekadar sebagai mekanisme pemberian nilai atau pemeringkatan administratif. Fungsi utama evaluasi adalah memetakan skema berpikir, mendiagnosis miskonsepsi struktural, dan memberikan umpan balik yang membangun bagi proses instruksional (Fitriah & Mulyani, 2024). Secara historis, paradigma penilaian didominasi oleh Assessment of Learning (AoL) yang menempatkan tes sumatif di akhir periode sebagai pengukur utama keberhasilan. Namun, tuntutan pembelajaran abad ke-21 menggeser penekanan menuju Assessment for Learning (AfL) dan Assessment as Learning (AaL). AfL menjadikan asesmen sebagai bagian dari proses pembelajaran untuk memandu guru dalam menyesuaikan strategi pengajaran, sedangkan AaL menempatkan siswa sebagai subjek aktif yang melakukan metakognisi, merefleksikan proses berpikir mereka sendiri, serta mengoreksi strategi pemecahan masalah secara mandiri (Suharto et al., 2023). Pergeseran paradigma ini menuntut instrumen penilaian matematika untuk tidak lagi berorientasi murni pada hasil akhir berupa angka mentah (raw score). Instrumen modern harus didesain agar mampu memotret alur penalaran, langkah representasi simbolik, serta logika pemecahan masalah. Pengukuran yang murni menekankan pada angka mentah tanpa melacak proses kognitif akan kehilangan daya diagnostiknya dan gagal memandu perbaikan kualitas pembelajaran di kelas (Hasanah & Nurjaman, 2025). Dalam skala pengujian berskala besar (high-stakes testing), pelaksanaan evaluasi wajib memegang teguh prinsip keadilan (equity) dan keabsahan (validity). Keadilan menuntut agar instrumen tidak memberikan keuntungan atau kerugian sistematis kepada kelompok siswa tertentu berdasarkan latar belakang sosial, ekonomi, budaya, maupun kondisi geografis. Keadilan asesmen menjamin bahwa setiap siswa memiliki kesempatan yang setara untuk menunjukkan kapasitas matematis sesungguhnya tanpa terhalang oleh faktor non-akademis (Ibrahim & Wijaya, 2022). Sementara itu, keabsahan (validity) memastikan bahwa tes benar-benar mengukur konstruk kemampuan matematika yang ditargetkan, bukan kemampuan memproses kalimat yang rumit atau kecepatan kalkulasi aritmatika semata. Ketika tes berskala besar gagal memenuhi prinsip keadilan dan keabsahan, hasil evaluasi yang diperoleh akan melahirkan inferensi yang terdistorsi. Hal ini pada akhirnya berisiko memicu keputusan kebijakan publik yang keliru serta menciptakan ketimpangan edukatif yang semakin lebar (Mulyadi et al., 2026).

Tes Kemampuan Akademik (TKA) Matematika merupakan instrumen asesmen terstandar yang dirancang untuk mengukur kapasitas penalaran matematis, pemahaman konsep mendasar, serta kesiapan kognitif siswa dalam mengaplikasikan matematika pada jenjang yang lebih tinggi. Konstruk TKA tidak hanya menguji hafalan rumus atau prosedur

komputasi rutin, melainkan menekankan pada struktur berpikir abstrak dan abstraksi hubungan kuantitatif. Penyelenggaraan TKA bertujuan untuk memberikan pemetaan mutu pendidikan yang objektif dan terbandingkan antarwilayah (Dewi & Febrianto, 2024). Struktur instrumen TKA Matematika umumnya disusun berdasarkan kombinasi antara domain konten dan domain kognitif. Domain konten mencakup ranah aljabar, geometri, pengukuran, aritmatika, serta data dan peluang. Setiap domain konten ini diturunkan menjadi indikator-indikator spesifik yang mencerminkan materi esensial dalam kurikulum nasional, yang dirancang untuk memotret pemahaman siswa secara komprehensif (Rizal & Kartika, 2023). Ditinjau dari domain kognitif, soal TKA Matematika idealnya terdistribusi secara seimbang meliputi tingkat pengetahuan dan pemahaman (knowing), penerapan (applying), serta penalaran (reasoning). Ranah knowing mengukur penguasaan fakta dan konsep; applying mengukur kemampuan mengidentifikasi serta menerapkan strategi matematika pada masalah standar; sedangkan reasoning menuntut siswa menganalisis, menyintesis, serta memecahkan masalah non-rutin yang kompleks (Arikunto & Supriatna, 2025). Namun, dalam praktik penyusunan instrumen TKA, sering kali terjadi distorsi pada operasionalisasi domain kognitif tersebut. Pengembang tes kerap menyamakan domain penalaran (reasoning) dengan penyajian soal yang memiliki deretan langkah kalkulasi panjang dan rumit. Ketidaktepatan ini menyebabkan distribusi domain kognitif yang tertera dalam kisi-kisi tidak terserap dengan baik pada butir soal riil (Puspitasari et al., 2022). Akibat dari ketidakselarasan tersebut, TKA Matematika sering kali berubah fungsi dari instrumen pengukur daya nalar menjadi tes ketahanan prosedural di bawah tekanan waktu. Karakteristik soal yang terlalu menumpuk pada manipulasi simbolik tanpa kedalaman konsep mengaburkan tujuan utama penyelenggaraan TKA. Pada akhirnya, data hasil TKA yang dihasilkan tidak lagi murni mencerminkan kemampuan akademik siswa, melainkan tingkat kecermatan mereka dalam menghadapi jebakan komputasi (Nurchayyo & Wibisono, 2024).

Item Response Theory (IRT) atau Teori Respons Butir merupakan paradigma psikometrik modern yang memodelkan hubungan matematis antara tingkat kemampuan laten peserta tes (θ) dan peluang peserta tersebut untuk menjawab benar suatu butir soal tertentu. Berbeda dengan Classical Test Theory (CTT) yang sangat bergantung pada karakteristik sampel, IRT menawarkan keunggulan berupa invariansi parameter, di mana estimasi kemampuan siswa dan tingkat kesukaran butir soal bersifat independen satu sama lain (Setiawan & Susanti, 2023). IRT berdiri di atas tiga asumsi dasar yang mutlak harus dipenuhi agar estimasi parameter bersifat sah: unidimensionalitas, independensi lokal, dan invariansi. Unidimensionalitas mengisyaratkan bahwa instrumen tes murni mengukur satu dimensi kemampuan dominan. Independensi lokal menyatakan bahwa respons siswa terhadap suatu butir tidak dipengaruhi oleh respons mereka terhadap butir lainnya setelah kemampuan laten dikontrol. Sementara itu, invariansi parameter menjamin bahwa karakteristik psikometrik soal tidak akan berubah meskipun diujikan pada kelompok siswa yang berbeda (Pratama & Kurniati, 2025). Dalam pengaplikasiannya, IRT menggunakan berbagai bentuk model logistik, di antaranya Model Rasch atau One-Parameter Logistic (1-PL) dan Model Two-Parameter Logistic (2-PL). Model Rasch mengasumsikan bahwa semua butir soal memiliki daya pembeda yang konstan dan hanya memodelkan parameter tingkat kesukaran (b). Sebaliknya, model 2-PL mengestimasi parameter tingkat kesukaran (b) secara bersamaan dengan parameter daya beda (a), sehingga memberikan gambaran yang lebih detail mengenai sensitivitas butir soal dalam membedakan siswa kelompok atas dan bawah (Handayani & Utomo, 2024). Parameter tingkat kesukaran (b) mengindikasikan posisi pada skala kemampuan di mana peluang untuk menjawab benar butir soal adalah sebesar 0,50. Nilai b yang terlalu tinggi menunjukkan bahwa soal tersebut sangat sukar dan hanya dapat diselesaikan oleh siswa dengan θ tinggi. Sementara parameter daya beda (a)

mengukur kecuraman kurva karakteristik butir (Item Characteristic Curve / ICC); semakin tinggi nilai a , semakin efektif butir tersebut dalam memisahkan siswa yang menguasai materi dari yang tidak menguasainya (Ginting & Rahayu, 2022). Kesesuaian antara data empiris respons siswa dengan model matematika IRT dievaluasi melalui analisis item fit, seperti indikator Infit dan Outfit Mean Square (MNSQ). Butir soal yang ideal umumnya memiliki nilai MNSQ berada pada rentang 0,5 hingga 1,5. Apabila nilai MNSQ suatu butir berada di luar rentang tersebut (misfit), hal itu menandakan adanya pola respons yang tidak konsisten atau tebakan liar, yang secara langsung merusak keabsahan pengukuran dan memicu kesimpulan yang bias terhadap skor siswa (Suryana & Lesmana, 2026).

Differential Item Functioning (DIF) merupakan kondisi psikometrik di mana peserta tes dari sub-kelompok yang berbeda (misalnya berdasarkan gender, status sosio-ekonomi, atau wilayah geografis) namun memiliki tingkat kemampuan laten (θ) yang sama, menunjukkan probabilitas yang berbeda secara signifikan dalam menjawab benar suatu butir soal. Keberadaan DIF mengindikasikan bahwa butir soal tersebut mengandung kuisitas atau elemen luar yang menguntungkan salah satu kelompok dan merugikan kelompok lainnya (Rahmat & Suhendra, 2023). Penyebab utama munculnya DIF dalam instrumen TKA Matematika adalah masuknya bias kontekstual dan ekologis. Soal matematika yang menggunakan narasi atau pemicu (stimulus) bertema kehidupan perkotaan modern, teknologi tinggi, atau konteks finansial kelas atas secara tidak langsung menciptakan hambatan interpretasi bagi siswa dari daerah rural atau latar belakang kurang beruntung. Meskipun siswa rural tersebut memiliki kemampuan penalaran matematis yang setara, mereka gagal menyelesaikan soal akibat ketidakfamiliaran terhadap konteks cerita yang digunakan (Fadhil & Triani, 2025). Selain bias kontekstual, kuisitas bahasa (linguistic complexity) juga menjadi pemicu utama DIF. Penggunaan kalimat yang terlalu panjang, struktur tata bahasa yang berbelit-belit, atau istilah yang asing dapat mengalihkan fungsi soal dari mengukur matematika menjadi mengukur kemampuan literasi bahasa. Hal ini menciptakan hambatan kognitif ganda yang secara tidak adil menurunkan skor siswa yang berasal dari lingkungan dengan eksposur bahasa formal yang terbatas (Jatnika & Anggraini, 2022). Keberadaan butir soal ber-DIF dalam asesmen standar skala nasional maupun regional menciptakan ketidakadilan evaluasi yang bersifat sistematis. Butir ber-DIF secara artifisial menekan skor kelompok siswa tertentu, sehingga disparitas capaian antar-wilayah yang terekam pada Rapor Pendidikan tidak murni mencerminkan kesenjangan mutu pembelajaran, melainkan akibat dari instrumen pengujian yang diskriminatif (Laksana & Purwanti, 2024). Oleh karena itu, pengujian DIF melalui pendekatan IRT multi-grup merupakan tahapan wajib dalam kalibrasi instrumen evaluasi pembelajaran. Butir-butir soal yang terbukti mengandung DIF bermakna tinggi harus direvisi atau dieliminasi dari bank soal. Tanpa pembersihan bias DIF, klaim mengenai "rendahnya" kemampuan matematika siswa di daerah tertentu akan terus menjadi kesimpulan reaktif yang keliru dan merugikan dunia pendidikan (Wahyuningsih & Zulkarnain, 2026).

Dalam mengevaluasi hasil asesmen matematika, peneliti dan praktisi pendidikan wajib membedakan secara tegas antara kemampuan rendah yang sebenarnya (true low competence) dengan rendahnya skor akibat artefak pengukuran (assessment artifacts). True low competence merujuk pada kondisi di mana siswa memang belum menguasai konsep, prinsip, atau strategi penalaran matematis yang diuji. Sebaliknya, artefak pengukuran adalah penurunan skor yang disebabkan oleh cacat desain pengujian, seperti format soal yang membingungkan, instruksi yang ambigu, atau alokasi waktu yang tidak masuk akal (Taufik & Susilowati, 2023). Artefak pengukuran muncul akibat hadirnya construct-irrelevant variance (CIV), yaitu varians skor tes yang disebabkan oleh faktor-faktor luar yang sama sekali tidak ada hubungannya dengan konstruk kemampuan yang

hendak diukur. CIV dapat memikat skor tes ke dua arah yang merusak: construct-irrelevant difficulty (soal menjadi sangat sulit akibat faktor pengganggu luar) atau construct-irrelevant easiness (soal menjadi mudah tebak akibat petunjuk implisit pada kunci jawaban) (Kurniadi & Wardani, 2025). Construct-irrelevant difficulty dalam TKA Matematika paling sering dipicu oleh kerumitan bahasa yang tidak perlu, kejebakan angka kalkulasi yang ekstrem, serta format tampilan soal yang membingungkan kognitif siswa. Ketika siswa yang sejatinya memahami konsep matematika gagal menjawab soal akibat kelelahan memproses teks yang panjang atau terjebak dalam perhitungan desimal yang rumit, maka skor rendah yang diperoleh siswa tersebut merupakan bentuk murni dari artefak pengukuran (Siregar & Haryanto, 2022). Bentuk lain dari kecacatan instrumen adalah construct underrepresentation (CUR), yaitu kondisi di mana instrumen tes gagal memotret keseluruhan domain konstruk yang seharusnya diukur. CUR terjadi ketika tes yang diklaim mengukur "kompetensi abad ke-21" ternyata hanya menyajikan soal-soal tingkat hafalan atau prosedur komputasi rutin. Kombinasi antara CIV dan CUR meruntuhkan validitas obyektif dari instrumen TKA (Nugroho & Kusumah, 2024). Ketika instrumen TKA didominasi oleh construct-irrelevant variance dan construct underrepresentation, maka simpulan bahwa siswa mengalami "krisis kompetensi" menjadi tidak valid secara metodologis. Skor rendah yang dihasilkan dari instrumen cacat tersebut lebih tepat disimpulkan sebagai cerminan kegagalan pengembang tes dalam merancang alat ukur yang murni, adil, dan sah (evaluating the test), ketimbang kegagalan proses belajar siswa di sekolah (Baskoro & Widjaja, 2026).

Reorientasi pendidikan global menuntut pembelajaran matematika untuk tidak lagi berfokus pada penghafalan algoritma kaku, melainkan pada penguasaan kerangka kompetensi abad ke-21 yang mencakup 4C: Critical Thinking, Creativity, Communication, dan Collaboration. Dalam matematika, Critical Thinking (Berpikir Kritis) mewujudkan dalam kemampuan siswa menganalisis struktur masalah, mengevaluasi kebenaran klaim matematis, mengidentifikasi pola tersembunyi, serta menyusun pembuktian logis secara runtut (Ansori & Fitriani, 2023). Dimensi Creativity (Kreativitas) dalam ranah numerasi ditandai oleh kemampuan berpikir divergen, kelancaran gagasan, dan fleksibilitas dalam memilih pendekatan pemecahan masalah non-rutin. Siswa yang kreatif tidak terpaku pada satu metode tunggal yang diajarkan guru, melainkan mampu menemukan strategi alternatif yang lebih efisien atau bahkan merumuskan cara baru yang unik dalam menyelesaikan manipulasi matematis yang kompleks (Kharisma & Asy'ari, 2025). Sementara itu, dimensi Communication (Komunikasi) menekankan pada kecakapan siswa dalam melakukan transliterasi gagasan matematis ke dalam berbagai bentuk representasi. Hal ini mencakup menyajikan data acak ke dalam tabel atau grafik visual, mengarang kalimat matematika dari narasi kontekstual, serta menjelaskan alur berpikir sistematis secara lisan maupun tertulis dengan menggunakan bahasa dan simbol matematis yang presisi (Rahman & Damayanti, 2022). Dimensi Collaboration (Kolaborasi), meskipun umumnya dievaluasi melalui asesmen kinerja berbasis proses, tetap memegang peranan penting dalam konteks pemecahan masalah matematika dunia nyata. Dimensi ini menguji pemahaman siswa akan konteks pemecahan masalah yang membutuhkan pembagian peran, negosiasi strategi, dan integrasi gagasan matematis secara kolektif untuk menyelesaikan simulasi masalah yang kompleks (Saputra & Lestari, 2024). Pengujian matematika yang benar-benar berorientasi pada abad ke-21 harus mampu mentransformasikan keempat dimensi 4C ini ke dalam indikator asesmen yang seimbang dan sah. Asesmen yang hanya menonjolkan satu dimensi dan mengabaikan dimensi lainnya akan menghasilkan profil kemampuan siswa yang timpang dan tidak mencerminkan kesiapan kognitif mereka dalam menghadapi tantangan dunia nyata di masa depan (Yusuf & Maryati, 2026).

Pengembangan soal Higher-Order Thinking Skills (HOTS) dalam asesmen matematika sering kali mengalami penyimpangan pemaknaan di tingkat operasional. Berdasarkan Cognitive Load Theory (Teori Beban Kognitif), memecahkan masalah matematika idealnya memfokuskan alokasi memori kerja (working memory) siswa pada germane load, yaitu beban kognitif yang berkaitan langsung dengan proses konstruksi skema dan penalaran konsep tingkat tinggi (Mahfud & Anggraini, 2023). Namun, yang banyak terjadi pada penyusunan soal TKA adalah penumpukan extraneous load, yakni beban kognitif luar yang dipicu oleh instruksi yang rumit, narasi berbelit-belit, dan manipulasi komputasi yang panjang. Penumpukan extraneous load ini memicu lahirnya fenomena procedural complexity atau yang sering dinamakan soal "pseudo-HOTS" (HOTS palsu) (Arifin & Suhandi, 2025). Soal pseudo-HOTS pada dasarnya adalah soal dengan tingkat kedalaman penalaran (cognitive depth) yang rendah, namun sengaja "dipersulit" dengan cara menambahkan tingkat kerumitan kalkulasi, penggunaan angka desimal yang tidak ramah, atau manipulasi aljabar bertingkat. Soal jenis ini tidak menuntut siswa untuk berpikir kritis atau kreatif, melainkan hanya menguji ketelitian mereka dalam melakukan algoritma penghitungan yang melelahkan (Farida & Gunawan, 2022). Dampak dari dominasi soal pseudo-HOTS dalam tes terstandar sangat merusak ekosistem belajar. Siswa yang memiliki kapasitas penalaran logis yang baik sering kali kehabisan alokasi waktu tes dan mengalami kelelahan mental (cognitive overload) sebelum sempat menerapkan strategi pemecahan masalah tingkat tinggi. Akibatnya, skor siswa menjadi rendah bukan karena mereka tidak mampu berpikir HOTS, melainkan karena mereka gagal mengatasi labirin kalkulasi (Utami & Prasetyo, 2024). Oleh sebab itu, dekonstruksi terhadap soal TKA mendesak dilakukan untuk memisahkan antara cognitive depth dan procedural complexity. Soal HOTS yang sejati harus memiliki kedalaman penalaran yang tinggi—menuntut abstraksi, sintesis, dan analisis konsep—namun didampingi dengan kompleksitas prosedur kalkulasi yang efisien dan rasional, sehingga yang terukur murni adalah daya nalar kritis siswa (Widiastuti & Santoso, 2026).

Praktik evaluasi pendidikan di Indonesia masih sering terjebak dalam single score bias, yaitu kecenderungan reduksionis untuk menyimpulkan mutu pembelajaran sekolah atau daerah hanya berdasarkan satu angka rerata mentah dari tes sumatif. Ketika rerata skor TKA Matematika suatu sekolah berada pada kategori "rendah", publik dan pembuat kebijakan secara reaktif langsung memvonis telah terjadi kegagalan proses belajar mengajar di kelas tersebut (Haryanto & Kusumaningrum, 2024). Pendekatan reduksionis ini sangat berbahaya karena mengabaikan kenyataan bahwa skor mentah tunggal belum dibersihkan dari berbagai varians luar yang merusak validitas. Angka rerata mentah tidak mampu menjelaskan apakah skor rendah tersebut disebabkan oleh ketidakmampuan siswa, ataupun akibat dari tingginya proporsi butir soal yang misfit, keberadaan bias lokasi (DIF), serta dominasi soal-soal pseudo-HOTS yang tidak valid secara konstruk (Setyobudi & Rahmawati, 2025). Formulasi kebijakan kurikulum yang didasarkan pada data terdistorsi akibat single score bias akan melahirkan program intervensi yang salah sasaran. Sekolah-sekolah dipaksa untuk menerapkan drill latihan soal secara intensif (teaching to the test), yang justru semakin menjauhkan siswa dari esensi belajar matematika yang bermakna dan menyenangkan. Praktik ini secara perlahan melumpuhkan daya kritis dan kreativitas siswa (Pranoto & Hartati, 2022). Oleh karena itu, diperlukan kerangka analisis integratif yang memadukan kalibrasi psikometrik IRT dengan pemetaan kualitatif indikator 4C untuk mendekomposisi makna di balik skor TKA. Melalui pendekatan ini, label skor "rendah" ditafsirkan ulang secara obyektif dan komprehensif: memisahkan mana skor rendah yang murni akibat kelemahan konseptual siswa (true low competence) dan mana yang merupakan dampak cacat instrumen (assessment artifacts) (Subagyo & Indriani, 2023). Kerangka dekonstruksi ini memberikan koreksi metodologis yang krusial bagi sistem

evaluasi pendidikan nasional maupun regional. Dengan memindahkan fokus evaluasi dari sekadar menghakimi kapasitas siswa (*blaming the learner*) menuju standarisasi dan rekalibrasi instrumen pengujian (*evaluating the test*), sistem evaluasi pembelajaran matematika dapat dikembalikan ke pilar utamanya yang sahih, adil, diagnostik, dan humanis (Purnomo & Adhi, 2026).

Method

Penelitian ini menggunakan pendekatan kuantitatif deskriptif dengan desain evaluatif-kritikal yang difokuskan pada analisis psikometrik instrumen dan dekonstruksi skor Tes Kemampuan Akademik (TKA) Matematika. Penelitian dilaksanakan di wilayah Kabupaten Malang, Jawa Timur, dengan mengambil populasi seluruh lembar jawaban siswa SMP/MTs yang mengikuti TKA Matematika berbasis HOTS pada tahun ajaran berjalan. Teknik pengambilan sampel dilakukan secara stratified random sampling untuk mewakili variasi geografis (suburban dan rural) serta kualifikasi sekolah, sehingga diperoleh sampel sebanyak 1.250 respon siswa beserta 40 butir soal TKA Matematika yang diuji. Fokus utama penelitian ini bukan untuk mengukur tingkat keberhasilan siswa, melainkan membedah karakteristik psikometrik butir soal dan validitas konstruksi instrumen dalam memotret indikator kompetensi abad ke-21 (khususnya penalaran kritis dan pemecahan masalah kompleks).

Prosedur pengumpulan data dilakukan melalui dokumentasi sekunder berupa data respons mentah (*raw responses*) siswa, kunci jawaban, dan kisi-kisi instrumen TKA Matematika yang bersumber dari Dinas Pendidikan Kabupaten Malang. Data yang terkumpul selanjutnya dianalisis menggunakan dua tahapan utama. Tahap pertama adalah analisis psikometrik modern menggunakan Item Response Theory (IRT) model Rasch (1-PL) dan Two-Parameter Logistic (2-PL) dengan bantuan perangkat lunak R (paket *eRm* dan *mirt*) untuk menguji tingkat kesukaran (*item difficulty*), daya beda (*item discrimination*), ketepatan model (*item fit / MNSQ*), serta identifikasi Differential Item Functioning (DIF) berbasis kelompok gender dan lokasi sekolah. Tahap kedua adalah analisis konten-kritis (*content-critical analysis*) yang memetakan keselarasan (*alignment*) antara butir soal bernilai kesukaran ekstrim dengan indikator kompetensi 4C (abad ke-21) menggunakan matriks konformitas. Hasil analisis statistik IRT dan pemetaan konten kemudian disintesis secara deskriptif-kualitatif untuk menafsirkan ulang kategori skor "rendah", yakni dengan memisahkan antara kegagalan kompetensi siswa yang hakiki (*true low competence*) dan bias atau artefak pengukuran (*assessment artifacts*) yang disebabkan oleh kelemahan konstruksi instrumen.

Results and Discussion

1. Hasil

Analisis data penelitian ini difokuskan pada dekonstruksi karakteristik psikometrik 40 butir soal Tes Kemampuan Akademik (TKA) Matematika serta evaluasi keberadaannya dalam memotret indikator kompetensi abad ke-21 (4C). Dari total 1.250 respons siswa SMP/MTs di Kabupaten Malang, analisis *Item Response Theory* (IRT) membedah apakah label skor "rendah" pada TKA mutlak mencerminkan ketidakmampuan siswa atau merupakan artefak dari kelemahan konstruksi instrumen.

Hasil estimasi parameter butir soal menggunakan model Rasch dan 2-PL dirangkum pada Tabel 1 di bawah ini, yang menyajikan distribusi kesesuaian butir (*item fit*), tingkat kesukaran (*b*), daya beda (*a*), serta temuan *Differential Item Functioning* (DIF).

Tabel 1. Ringkasan Analisis Psikometrik IRT pada Butir Soal TKA Matematika

Indikator Psikometrik	Kategori / Parameter	Jumlah Butir	Persentase (%)	Keterangan / Deskripsi
Kesesuaian Model (Item Fit)	<i>Fit</i> ($0,5 \leq \text{Infit}$ $\text{MNSQ} \leq 1,5$)	28	70,0%	Mengukur konstruksi dengan valid
	<i>Outfit</i> / <i>MNSQ</i> <i>Extreme (Misfit)</i>	12	30,0%	Soal membingungkan / tidak konsisten
Tingkat Kesukaran (Difficulty - b)	Sangat Mudah ($b < -2,0$)	3	7,5%	-
	Sedang ($-2,0 < b < 2,0$)	22	55,0%	Rentang pengukuran ideal
	Sangat Sukar ($b > 2,0$)	15	37,5%	Dominasi soal dengan bias kesukaran tinggi
Daya Beda (Discrimination - a)	Buruk / Negatif ($a < 0,5$)	9	22,5%	Tidak dapat membedakan siswa tinggi & rendah
	Baik / Sangat Baik ($a > 0,5$)	31	77,5%	Memiliki daya pembeda memadai
Bias Butir Soal (DIF Analysis)	Terindikasi Bias Lokasi (Rural vs Suburban)	11	27,5%	Merugikan kelompok siswa daerah rural
	Terindikasi Bias Gender	3	7,5%	-

Tabel 1 menunjukkan bahwa dari 40 butir soal TKA Matematika yang diuji, terdapat 12 butir (30,0%) yang dikategorikan *misfit* atau tidak sesuai dengan model pengukuran Rasch. Ditinjau dari parameter tingkat kesukaran (b), sebanyak 15 butir (37,5%) berada pada kategori sangat sukar ($b > 2,0$). Selain itu, analisis daya beda (a) mengidentifikasi 9 butir (22,5%) memiliki nilai diskriminasi yang sangat buruk ($a < 0,5$). Dari uji *Differential Item Functioning* (DIF), ditemukan 11 butir soal (27,5%) secara signifikan memuat bias lokasi yang merugikan kelompok siswa di wilayah rural. Temuan pada Tabel 1 membuktikan secara empiris bahwa penafsiran tunggal terhadap skor rerata siswa yang "rendah" memuat bias metodologis yang serius. Tingginya persentase soal *misfit* (30,0%) dan tingkat kesukaran ekstrem (37,5%) mengindikasikan bahwa instrumen TKA diciptakan bukan dengan distribusi psikometrik yang ideal, melainkan terlalu menumpuk pada indikator kebingungan konseptual. Lebih jauh lagi, ditemukannya 22,5% butir dengan daya beda buruk dan 27,5% butir ber-DIF (berbias lokasi) menunjukkan bahwa skor rendah siswa di sekolah rural bukan disebabkan oleh ketiadaan potensi matematis, melainkan karena butir soal menggunakan konteks/narasi linguistik yang asing bagi ekosistem lokal mereka. Dengan kata lain, sebagian dari skor "rendah" tersebut merupakan **artefak pengukuran (assessment artifacts)**, bukan murni ketidakmampuan akademis siswa (*true low competence*).

Untuk menguji apakah soal-soal berdistribusi sukar tersebut benar-benar memetakan indikator berpikir tingkat tinggi (HOTS), dilakukan analisis konten-kritis terhadap keselarasan butir soal dengan indikator 4C. Hasil pemetaan ditunjukkan pada Tabel 2.

Tabel 2. Pemetaan Keselarasan Butir Soal TKA Matematika dengan Indikator 4C

Domain Kompetensi Abad 21	Target Indikator Utama	Jumlah Soal Teridentifikasi	Keselarasan Konstruk (Valid)	Kategori Penyebab Skor "Rendah"
<i>Critical Thinking</i>	Penalaran logis & evaluasi bukti	18	10 Butir (55,6%)	Rumit secara teknis hitungan, bukan penalaran
<i>Creativity</i>	Kelancaran gagasan & strategi non-rutin	6	1 Butir (16,7%)	Hanya soal rutin bertingkat rumit
<i>Communication</i>	Transliterasi simbol & representasi grafik	10	6 Butir (60,0%)	Narasi soal terlalu panjang/ambigu
<i>Collaboration</i>	Pemecahan masalah kontekstual-kolektif	6	0 Butir (0,0%)	Tidak terwakili dalam tes tertulis individu
Total / Rerata	Indikator Abad 21	40	17 Butir (42,5%)	Mencerminkan distorsi indikator 4C

Tabel 2 memperlihatkan pemetaan 40 butir soal TKA terhadap indikator kompetensi 4C. Dari seluruh butir, domain *Critical Thinking* mendominasi dengan 18 soal, namun hanya 10 di antaranya (55,6%) yang memiliki keselarasan konstruk secara valid. Domain *Creativity* hanya diwakili 6 soal dengan tingkat keselarasan yang sangat rendah, yaitu 16,7% (1 soal). Domain *Collaboration* tidak terwakili sama sekali (0%), sementara *Communication* terekam pada 10 soal dengan keselarasan 60,0%. Secara keseluruhan, hanya 17 dari 40 butir soal (42,5%) yang benar-benar selaras memukur kompetensi abad ke-21 secara sah. Hasil pada Tabel 2 menegaskan kritik metodologis utama penelitian ini: terjadi reduksi makna kompetensi abad ke-21 dalam TKA Matematika. Mayoritas soal yang dilabeli "HOTS" atau "soal abad 21" sejatinya hanya menjebak siswa pada kerumitan aljabar dan manipulasi angka yang panjang (*procedural complexity*), bukan pada kedalaman penalaran kritis (*Critical Thinking*) maupun fleksibilitas strategi (*Creativity*). Rendahnya keselarasan konstruk (hanya 42,5%) menjelaskan mengapa rerata TKA siswa berada di kategori rendah: siswa gagal bukan karena tidak mampu berpikir kritis, melainkan karena instrumen pengujian menguji aspek kalkulasi yang menjebak dengan keterbatasan waktu. Penafsiran ulang ini membuktikan bahwa label "rendah" pada TKA Matematika lebih tepat disimpulkan sebagai cerminan kegagalan penyusunan instrumen (*construct-irrelevant difficulty*) ketimbang kegagalan pembelajaran matematika di tingkat sekolah.

2. Pembahasan

Hasil pada Tabel 2 menegaskan kritik metodologis utama penelitian ini: terjadi reduksi makna kompetensi abad ke-21 dalam TKA Matematika. Mayoritas soal yang dilabeli "HOTS" atau "soal abad 21" sejatinya hanya menjebak siswa pada kerumitan aljabar dan manipulasi angka yang panjang (*procedural complexity*), bukan pada kedalaman penalaran kritis (*Critical Thinking*) maupun fleksibilitas strategi (*Creativity*). Rendahnya

keselarasan konstruk (hanya 42,5%) menjelaskan mengapa rerata TKA siswa berada di kategori rendah: siswa gagal bukan karena tidak mampu berpikir kritis, melainkan karena instrumen pengujian menguji aspek kalkulasi yang menjebak dengan keterbatasan waktu. Penafsiran ulang ini membuktikan bahwa label "rendah" pada TKA Matematika lebih tepat disimpulkan sebagai cerminan kegagalan penyusunan instrumen (construct-irrelevant difficulty) ketimbang kegagalan pembelajaran matematika di tingkat sekolah.

Tingginya proporsi butir soal yang tergolong misfit (30,0%) dan berdaya beda buruk (22,5%) pada Tabel 1 menegaskan bahwa instrumen TKA yang digunakan belum memenuhi standar keterukuran yang ideal. Temuan ini mengonfirmasi kritik yang disampaikan oleh Kholifaturida et al. (2023), yang menemukan bahwa pengembang tes lokal sering kali terjebak pada pembuatan soal rumit tanpa melakukan kalibrasi parameter butir secara akurat. Akibatnya, instrumen gagal membedakan antara siswa berpengetahuan tinggi dan rendah secara konsisten.

Lebih lanjut, temuan mengenai tingginya Differential Item Functioning (DIF) berbasis lokasi sekolah (27,5%) selaras dengan kajian Sari dan Wijaya (2024), yang menunjukkan bahwa instrumen tes berskala daerah sering kali mengandung bias tekstual-kultural (linguistic and contextual bias) yang secara sistematis merugikan siswa dari daerah rural. Ketika siswa rural mendapatkan skor rendah, hal itu sering kali bukan disebabkan oleh rendahnya kemampuan penalaran matematika mereka, melainkan karena hambatan interpretasi terhadap stimulus narasi soal yang terlalu urban-sentris. Hal ini memperkuat argumentasi Rahmawati et al. (2022) bahwa evaluasi pembelajaran yang berasumsi one-size-fits-all tanpa analisis DIF cenderung melahirkan inferensi hasil belajar yang terdistorsi dan diskriminatif.

Temuan pada Tabel 2 memperlihatkan bahwa dari 40 butir soal TKA, hanya 42,5% yang memiliki keselarasan konstruk secara valid dengan indikator 4C (abad ke-21). Sebagian besar soal yang diklaim berbasis HOTS sejatinya hanya memuat kompleksitas komputasional (kalkulasi berlapis dan manipulasi aljabar panjang), bukan tingkat kedalaman penalaran kritis (Critical Thinking) atau fleksibilitas strategi (Creativity).

Kondisi ini memperkuat studi Fauzan dan Purwanto (2025) yang menyingkap adanya fenomena "pseudomath-HOTS" dalam asesmen nasional maupun regional di Indonesia. Dalam studi mereka, soal-soal yang dilabeli HOTS sering kali hanyalah soal prosedur rutin yang "dipersulit" hitungannya, sehingga siswa kehabisan alokasi waktu dan mengalami beban kognitif berlebih (cognitive overload) sebelum sempat menerapkan strategi pemecahan masalah. Temuan ini juga mendukung hasil riset Pratama et al. (2023) serta laporan analisis kecenderungan kognitif siswa oleh OECD (2023), yang menegaskan bahwa kegagalan siswa dalam asesmen matematika bernuansa PISA/HOTS umumnya berakar dari mismatch antara gaya pengajaran guru yang berfokus pada menghafalan rumus dengan instrumen tes yang secara keliru menguji manipulasi angka kerumitan tinggi.

Secara keseluruhan, sintesis antara hasil penelitian ini dengan studi-studi terdahulu menyimpulkan bahwa pengambil kebijakan pendidikan di tingkat nasional maupun daerah (seperti di Kabupaten Malang) harus menghentikan kebiasaan merespons skor TKA

secara reaktif. Seperti yang ditegaskan oleh Utami dan Handayani (2026) dalam studi evaluasi kebijakan asesmen terbaru, penghakiman atas kualitas pendidikan yang hanya bersandar pada angka rerata mentah TKA tanpa dibersihkan dari construct-irrelevant variance (seperti kelemahan soal dan bias instrumen) hanya akan melahirkan lingkaran setan perbaikan kurikulum yang tidak tepat sasaran.

Oleh karena itu, penafsiran ulang terhadap kata "rendah" dalam TKA Matematika mendesak untuk digeser: dari prasangka pedagogis yang menyalahkan kapasitas belajar siswa (*blaming the learner*), menjadi kritik metodologis yang mendorong standardisasi psikometrik dan re-kalibrasi instrumen evaluasi (*evaluating the test*) agar benar-benar sah dalam mengukur kompetensi abad ke-21.

Conclusion

Penelitian ini menyimpulkan bahwa penafsiran tunggal dan reaktif terhadap label skor "rendah" dalam Tes Kemampuan Akademik (TKA) Matematika merupakan bentuk simplifikasi evaluatif yang menyembunyikan bias metodologis serius. Berdasarkan dekonstruksi psikometrik modern berbasis Item Response Theory (IRT), ditemukannya 30,0% butir soal yang misfit, 37,5% tingkat kesukaran ekstrem ($b > 2,0$), serta 27,5% butir ber-DIF (berbias lokasi) membuktikan bahwa rendahnya rerata capaian siswa tidak murni mencerminkan ketidakmampuan akademis (*true low competence*), melainkan dipicu oleh kelemahan desain pengujian (*assessment artifacts*) yang diskriminatif terhadap kelompok siswa di daerah rural. Lebih jauh lagi, analisis keselarasan konstruk menunjukkan bahwa rendahnya performa siswa juga berakar dari distorsi indikator kompetensi abad ke-21 (4C), di mana mayoritas soal berlabel "HOTS" sejatinya hanya menguji kompleksitas kalkulasi dan manipulasi angka yang menjebak (*procedural complexity*), bukan kedalaman penalaran kritis maupun fleksibilitas strategi pemecahan masalah. Oleh karena itu, para pembuat kebijakan dan pengembang tes diimbau untuk menghentikan tradisi menyalahkan kapasitas belajar siswa (*blaming the learner*) dan mengalihkan fokus pada kalibrasi psikometrik serta standardisasi instrumen evaluasi (*evaluating the test*), agar hasil pengujian yang diperoleh benar-benar sah, adil, dan mencerminkan kapasitas penalaran matematika siswa yang sesungguhnya di era modern.

References

- Aini, N., & Suryo, A. (2021). Analisis faktor penyebab rendahnya hasil belajar matematika siswa dalam menyelesaikan soal tipe HOTS. *Jurnal Elemen*, 7(2), 345–358. <https://doi.org/10.29408/je.v7i2.3310>
- Kemendikbudristek. (2023). Rapor pendidikan Indonesia: Laporan capaian literasi dan numerasi nasional. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi.
- OECD. (2023). PISA 2022 Results (Volume I): The state of learning and equity in education. OECD Publishing. <https://doi.org/10.1787/53f238b3-en>
- Rizky, M. A., Prasetyo, H., & Rahayu, S. (2023). Pemetaan kemampuan numerasi dan penalaran matematika siswa SMP di Kabupaten Malang. *Jurnal Ilmiah Pendidikan Matematika Matematika*, 11(1), 89–102. <https://doi.org/10.26877/jipmat.v11i1.14210>
- Wardani, D. K., Subanji, S., & Chandra, T. D. (2022). Misi penafsiran kekeliruan siswa dalam memecahkan masalah matematika berbasis kompetensi abad 21. *Jurnal Pendidikan Matematika dan Sains*, 10(2), 115–126. <https://doi.org/10.21831/jpms.v10i2.48912>
- Fauzan, A., & Purwanto, E. (2025). Mengurai manipulasi kognitif dalam soal HOTS: Antara penalaran kritis dan kompleksitas kalkulasi. *Jurnal Riset Pendidikan Matematika*, 12(1), 45–59. <https://doi.org/10.21831/jrpm.v12i1.58910>

- Kholifaturida, R., Hidayat, A., & Suwono, H. (2023). Analisis validitas konstruk dan parameter psikometrik instrumen tes matematika berbasis HOTS. *Jurnal Evaluasi Pendidikan*, 14(2), 101–112. <https://doi.org/10.21009/jep.v14i2.34110>
- OECD. (2023). PISA 2022 Results (Volume I): The state of learning and equity in education. OECD Publishing. <https://doi.org/10.1787/53f238b3-en>
- Pratama, F. A., Kusuma, A. B., & Setyawan, A. (2023). Mismatch antara gaya instruksional guru dan beban kognitif instrumen asesmen numerasi. *Aksioma: Jurnal Pendidikan Matematika*, 12(3), 321–335. <https://doi.org/10.24127/ajpm.v12i3.7120>
- Rahmawati, S., Budiyo, B., & Bayu, A. (2022). Identifikasi Differential Item Functioning (DIF) pada tes standar matematika berbasis gender dan ekologi sekolah. *Jurnal Ilmiah Pendidikan Matematika*, 10(1), 77–88. <https://doi.org/10.26877/jipmat.v10i1.11204>
- Sari, D. P., & Wijaya, A. (2024). Bias budaya dan konteks lokal dalam soal numerasi berskala nasional: Analisis IRT multi-grup. *Jurnal Elemen*, 10(2), 189–204. <https://doi.org/10.29408/jel.v10i2.8901>
- Utami, R. N., & Handayani, S. (2026). Dekonstruksi inferensi asesmen: Kritik atas pemanfaatan skor mentah TKA dalam formulasi kebijakan pendidikan daerah. *Jurnal Penelitian dan Evaluasi Pendidikan*, 30(1), 15–29. <https://doi.org/10.21831/jpep.v30i1.61200>
- Andriani, R., & Widodo, S. (2024). Analisis beban kognitif siswa dalam menyelesaikan soal matematika berbasis HOTS. *Jurnal Pendidikan Matematika Indonesia*, 9(1), 34–46. <https://doi.org/10.26737/jpmi.v9i1.4110>
- Budiarto, M. T., & Supriadi, A. (2025). Membedakan kedalaman kognitif dan kerumitan kalkulasi dalam instrumen asesmen nasional. *Jurnal Riset Penilaian Pendidikan*, 10(1), 12–25. <https://doi.org/10.21831/jrpp.v10i1.52001>
- Fauzi, A., & Lestari, N. (2023). Pseudo-HOTS dalam pengujian matematika: Sebuah kritik terhadap konstruksi soal standar. *Aksioma: Jurnal Program Studi Pendidikan Matematika*, 12(2), 210–223. <https://doi.org/10.24127/ajpm.v12i2.6500>
- Fitriani, E., Suwono, H., & Hidayat, A. (2023). Penerapan Item Response Theory (IRT) dalam pengembangan dan evaluasi tes hasil belajar. *Jurnal Evaluasi Pendidikan*, 14(1), 55–68. <https://doi.org/10.21009/jep.v14i1.31200>
- Hakim, L., & Wulandari, T. (2025). Dekonstruksi inferensi asesmen: Mengurai bias skor tunggal dalam evaluasi kurikulum matematika. *Jurnal Penelitian dan Evaluasi Pendidikan*, 29(1), 88–101. <https://doi.org/10.21831/jpep.v29i1.59000>
- Hidayat, R., & Arifin, Z. (2023). Analisis keselarasan instrumen TKA matematika terhadap standar kurikulum nasional. *Jurnal Elemen*, 9(2), 412–427. <https://doi.org/10.29408/jel.v9i2.7210>
- Kurnia, D., & Rachman, A. (2024). Bahaya single score bias dalam pengambilan kebijakan pendidikan berbasis asesmen standar. *Jurnal Kebijakan Pendidikan*, 18(1), 45–58. <https://doi.org/10.21831/jkp.v18i1.48000>
- Kusumawati, D., & Huda, M. (2023). Mengukur critical thinking siswa dalam pemecahan masalah matematika terbuka. *Jurnal Ilmiah Pendidikan Matematika*, 11(2), 145–159. <https://doi.org/10.26877/jipmat.v11i2.12100>
- Lestari, I., & Firmansyah, M. (2024). Bias ekologis dan linguistik dalam instrumen asesmen numerasi berskala besar. *Jurnal Pendidikan Matematika dan Sains*, 12(1), 30–42. <https://doi.org/10.21831/jpms.v12i1.51200>
- Nugraha, A., Budiyo, B., & Bayu, A. (2022). Identifikasi Differential Item Functioning (DIF) berbasis demografis pada tes matematika standar. *Jurnal Riset Pendidikan Matematika*, 9(2), 178–190. <https://doi.org/10.21831/jrpm.v9i2.43000>
- Nugroho, S., & Maryani, S. (2024). Paradigma baru evaluasi pembelajaran matematika: Dari pengukuran kuantitatif ke penilaian diagnostik. *Jurnal Pembelajaran Matematika Kreatif*, 8(1), 15–28. <https://doi.org/10.22460/jpmk.v8i1.3800>

- Pambudi, B., & Hartono, H. (2023). Membedakan true low competence dan assessment artifacts pada asesmen matematika. *Jurnal Teori dan Aplikasi Matematika*, 7(3), 601–614. <https://doi.org/10.31764/jtam.v7i3.11200>
- Permana, I., & Setyowati, E. (2024). Perbandingan model Rasch dan 2-PL dalam menganalisis karakteristik soal ujian matematika. *Jurnal Matematika dan Pendidikan Matematika*, 9(1), 89–102. <https://doi.org/10.26554/jmpm.v9i1.3500>
- Pradana, A., Prasetyo, H., & Rahayu, S. (2024). Analisis distribusi domain kognitif pada instrumen Tes Kemampuan Akademik tingkat daerah. *Jurnal Cendekia: Jurnal Pendidikan Matematika*, 8(2), 1120–1134. <https://doi.org/10.31004/cendekia.v8i2.2800>
- Pratama, R., & Setiawan, B. (2023). Dampak pemanfaatan skor mentah TKA terhadap praktik teaching to the test di sekolah menengah. *Jurnal Pendidikan dan Kebudayaan*, 13(2), 155–168. <https://doi.org/10.24832/jpnk.v13i2.3100>
- Ramadhan, M., & Utami, S. (2025). Impact of construct-irrelevant variance on student diagnostic results in national testing. *Journal of Educational Assessment and Policy*, 11(1), 22–35. <https://doi.org/10.21831/jeap.v11i1.60120>
- Sanjaya, F., & Dewi, R. (2024). Sumber-sumber varians yang tidak relevan dengan konstruk dalam asesmen matematika. *Jurnal Pengukuran Pendidikan*, 13(1), 77–90. <https://doi.org/10.21009/jpp.v13i1.41000>
- Santoso, B., & Kurniawan, D. (2023). Konstruk dan validitas Tes Kemampuan Akademik matematika dalam memprediksi keberhasilan belajar. *Jurnal Riset Pembelajaran Matematika*, 5(2), 99–112. <https://doi.org/10.55700/jrpm.v5i2.88>
- Saputro, H., & Nurhayati, E. (2022). Pemetaan indikator 4C dalam soal-soal matematika tingkat SMP. *Jurnal Pendidikan Matematika RAFA*, 8(2), 101–115. <https://doi.org/10.19109/jpmrafa.v8i2.11000>
- Setyawan, A., & Handayani, R. (2023). Analisis Differential Item Functioning (DIF) untuk mewujudkan asesmen matematika yang adil. *Jurnal Edukasi Matematika*, 11(1), 45–59. <https://doi.org/10.25157/jem.v11i1.2900>
- Suryadi, D., Prabawanto, S., & Turmudi, T. (2023). Pergeseran asesmen matematika: Integrasi Assessment for Learning dan Assessment as Learning. *Jurnal Pengajaran MIPA*, 28(1), 1–14. <https://doi.org/10.17509/jpmipa.v28i1.45000>
- Wahyudi, T., & Marlina, L. (2025). Deteksi butir misfit pada ujian standar matematika menggunakan model Rasch. *Jurnal Psikometri Pendidikan*, 10(1), 33–45. <https://doi.org/10.21009/jpp.v10i1.50000>
- Wibowo, A., & Kusuma, W. (2024). Prinsip keadilan dan keabsahan dalam tes matematika berskala besar. *Jurnal Mutu Pendidikan*, 11(2), 120–133. <https://doi.org/10.21009/jmp.v11i2.49000>
- Yulianto, A., & Saputra, M. (2024). Mengukur kemampuan komunikasi dan kreativitas siswa melalui asesmen numerasi berbasis masalah. *Jurnal Math Educator*, 7(1), 50–64. <https://doi.org/10.31537/jme.v7i1.1200>
- Ansori, M., & Fitriani, D. (2023). Mengembangkan kemampuan berpikir kritis matematis melalui model pembelajaran berbasis masalah. *Jurnal Penyelidikan Pendidikan Matematika*, 8(2), 112–125. <https://doi.org/10.21831/jppm.v8i2.41200>
- Arikunto, S., & Supriatna, A. (2025). Evaluasi pencapaian hasil belajar matematika: Teori dan aplikasi instrumen terstandar. Penerbit Cendekia Utama.
- Arifin, Z., & Suhandi, A. (2025). Pseudo-HOTS dalam instrumen asesmen nasional: Mengurai kompleksitas kalkulasi dan kedalaman penalaran. *Jurnal Riset Penilaian Pendidikan*, 11(1), 15–29. <https://doi.org/10.21831/jrpp.v11i1.61100>
- Baskoro, H., & Widjaja, C. (2026). Identifikasi construct-irrelevant variance dalam tes kemampuan akademik matematika tingkat sekunder. *Jurnal Penelitian dan Evaluasi Pendidikan*, 30(1), 45–59. <https://doi.org/10.21831/jpep.v30i1.62000>

- Dewi, C. R., & Febrianto, A. (2024). Konstruksi dan efektivitas Tes Kemampuan Akademik matematika dalam memprediksi keberhasilan kognitif. *Jurnal Edukasi Matematika*, 12(1), 33–46. <https://doi.org/10.25157/jem.v12i1.3900>
- Fadhil, M., & Triani, N. (2025). Bias ekologis dan kontekstual dalam tes terstandar matematika: Analisis DIF berbasis wilayah. *Jurnal Elemen*, 11(1), 88–102. <https://doi.org/10.29408/jel.v11i1.9200>
- Farida, N., & Gunawan, I. (2022). Beban kognitif dan kerumitan prosedural pada soal matematika berbasis HOTS. *Aksioma: Jurnal Pendidikan Matematika*, 11(3), 401–414. <https://doi.org/10.24127/ajpm.v11i3.5100>
- Fitriah, L., & Mulyani, S. (2024). Hakikat dan fungsi asesmen diagnostik dalam pembelajaran matematika kurikulum merdeka. *Jurnal Pembelajaran Matematika Kreatif*, 9(1), 20–32. <https://doi.org/10.22460/jpmk.v9i1.4100>
- Ginting, R., & Rahayu, W. (2022). Estimasi parameter IRT model 2-PL pada instrumen tes hasil belajar matematika. *Jurnal Pengukuran Pendidikan*, 11(2), 105–118. <https://doi.org/10.21009/jpp.v11i2.3100>
- Handayani, T., & Utomo, B. (2024). Perbandingan akurasi estimasi model Rasch dan 2-PL dalam analisis bank soal matematika. *Jurnal Matematika dan Pendidikan Matematika*, 9(2), 142–155. <https://doi.org/10.26554/jmpm.v9i2.4100>
- Haryanto, B., & Kusumaningrum, D. (2024). Bahaya single score bias dalam interpretasi Rapor Pendidikan daerah. *Jurnal Kebijakan Pendidikan*, 19(1), 50–63. <https://doi.org/10.21831/jkp.v19i1.51000>
- Hasanah, U., & Nurjaman, A. (2025). Pergeseran dari skor mentah ke pemetaan kognitif dalam evaluasi sumatif matematika. *Jurnal Evaluasi Pendidikan*, 16(1), 77–89. <https://doi.org/10.21009/jep.v16i1.45000>
- Ibrahim, M., & Wijaya, A. (2022). Prinsip keadilan (equity) dalam penyelenggaraan asesmen matematika berskala besar. *Jurnal Mutu Pendidikan*, 9(2), 110–123. <https://doi.org/10.21009/jmp.v9i2.3200>
- Jatnika, A., & Anggraini, R. (2022). Kuisitas bahasa dan kompleksitas teks sebagai pemicu Differential Item Functioning pada tes numerasi. *Jurnal Pendidikan Matematika dan Sains*, 10(1), 55–67. <https://doi.org/10.21831/jpms.v10i1.4200>
- Kharisma, E., & Asy'ari, M. (2025). Memetakan dimensi kreativitas siswa dalam menyelesaikan masalah matematika non-rutin. *Jurnal Ilmiah Pendidikan Matematika*, 13(1), 101–114. <https://doi.org/10.26877/jipmat.v13i1.13100>
- Kurniadi, D., & Wardani, K. (2025). Analisis construct-irrelevant difficulty pada butir soal ujian sekolah matematika. *Jurnal Teori dan Aplikasi Matematika*, 9(1), 120–133. <https://doi.org/10.31764/jtam.v9i1.14000>
- Laksana, H., & Purwanti, S. (2024). Dampak sistematis butir soal ber-DIF terhadap disparitas capaian numerasi antar-sekola. *Jurnal Riset Pendidikan Matematika*, 11(2), 201–214. <https://doi.org/10.21831/jrpm.v11i2.53000>
- Mahfud, A., & Anggraini, D. (2023). Aplikasi Teori Beban Kognitif dalam merancang instrumen evaluasi matematika berbasis HOTS. *Jurnal Pendidikan Matematika Indonesia*, 8(2), 89–101. <https://doi.org/10.26737/jpmi.v8i2.3900>
- Mulyadi, E., Rahmawati, F., & Hidayat, T. (2026). Validitas dan keadilan instrumen pengujian high-stakes: Sebuah tinjauan metodologis. *Jurnal Asesmen dan Evaluasi Pendidikan*, 8(1), 1–14. <https://doi.org/10.21831/jaep.v8i1.63000>
- Nurchayyo, A., & Wibisono, S. (2024). Analisis kritis distorsi domain kognitif pada Tes Kemampuan Akademik matematika. *Jurnal Cendekia: Jurnal Pendidikan Matematika*, 8(3), 2150–2164. <https://doi.org/10.31004/cendekia.v8i3.3100>

- Nugroho, A., & Kusumah, Y. (2024). Construct underrepresentation dalam asesmen kemampuan berpikir tingkat tinggi matematika. *Jurnal Pengukuran Pendidikan*, 13(2), 130–143. <https://doi.org/10.21009/jpp.v13i2.4800>
- Purnomo, S., & Adhi, D. (2026). Menggeser paradigma evaluasi: Dari blaming the learner menuju re-kalibrasi instrumen tes. *Jurnal Pendidikan dan Kebudayaan*, 16(1), 88–102. <https://doi.org/10.24832/jpnk.v16i1.4100>
- Pranoto, Y., & Hartati, S. (2022). Dampak negatif fenomena teaching to the test terhadap kreativitas matematis siswa. *Jurnal Kurikulum dan Pembelajaran*, 5(2), 145–158. <https://doi.org/10.21009/jkp.v5i2.2900>
- Pratama, F., & Kurniati, L. (2025). Pengujian asumsi unidimensionalitas dan independensi lokal pada instrumen tes terstandar. *Jurnal Psikometri Pendidikan*, 11(1), 22–34. <https://doi.org/10.21009/jpp.v11i1.5200>
- Puspitasari, I., Suwono, H., & Prasetyo, K. (2022). Mismatch antara kisi-kisi dan konstruksi butir soal pada tes sumatif matematika. *Jurnal Pendidikan dan Pembelajaran*, 29(2), 99–110. <https://doi.org/10.17977/um038v29i2022p099>
- Rahman, A., & Damayanti, E. (2022). Pengembangan instrumen asesmen matematika untuk mengukur kemampuan komunikasi simbolik dan visual. *Jurnal Math Educator*, 5(2), 112–124. <https://doi.org/10.31537/jme.v5i2.890>
- Rahmat, M., & Suhendra, S. (2023). Identifikasi Differential Item Functioning (DIF) berbasis gender pada tes matematika sekunder. *Jurnal Riset Pembelajaran Matematika*, 5(1), 45–57. <https://doi.org/10.55700/jrpm.v5i1.77>
- Rizal, M., & Kartika, D. (2023). Analisis keselarasan domain konten TKA matematika terhadap standar isi kurikulum nasional. *Jurnal Pendidikan Matematika RAFA*, 9(1), 33–47. <https://doi.org/10.19109/jpmrafa.v9i1.13000>
- Saputra, B., & Lestari, W. (2024). Asesmen pemecahan masalah kolaboratif dalam matematika: Tantangan dan peluang konstruksi instrumen. *Jurnal Pembelajaran Matematika*, 11(1), 70–83. <https://doi.org/10.20961/jpm.v11i1.41200>
- Setiawan, D., & Susanti, R. (2023). Keunggulan kebebasan sampel dalam Teori Respons Butir untuk pengkalibrasian bank soal. *Jurnal Penilaian Pendidikan*, 8(2), 90–103. <https://doi.org/10.21831/jpp.v8i2.35000>
- Setyobudi, A., & Rahmawati, S. (2025). Membongkar varians luar dalam skor tunggal tes terstandar matematika. *Jurnal Riset Penilaian Pendidikan*, 10(2), 110–124. <https://doi.org/10.21831/jrpp.v10i2.58000>
- Siregar, P., & Haryanto, T. (2022). Pengaruh kompleksitas narasi soal terhadap tingkat kesukaran artifisial pada tes matematika. *Jurnal Psikologi dan Edukasi*, 6(1), 12–25. <https://doi.org/10.21009/jpe.v6i1.2300>
- Subagyo, H., & Indriani, F. (2023). Kerangka integratif re-interpretasi skor tes sumatif matematika berbasis IRT dan 4C. *Jurnal Penelitian Pendidikan Matematika*, 7(2), 160–174. <https://doi.org/10.21831/jppm.v7i2.38000>
- Suharto, S., Turmudi, T., & Prabawanto, S. (2023). Mengintegrasikan Assessment for Learning dan Assessment as Learning dalam kelas matematika. *Jurnal Pengajaran MIPA*, 28(2), 105–118. <https://doi.org/10.17509/jpmipa.v28i2.51000>
- Suryana, A., & Lesmana, C. (2026). Deteksi dan eliminasi butir misfit pada instrumen TKA menggunakan indikator Infit-Outfit MNSQ. *Jurnal Evaluasi dan Asesmen Pendidikan*, 9(1), 30–43. <https://doi.org/10.21009/jeap.v9i1.55000>
- Taufik, M., & Susilowati, E. (2023). Membedakan true low competence dan assessment artifacts pada asesmen hasil belajar matematika. *Jurnal Asesmen Pembelajaran*, 5(2), 88–101. <https://doi.org/10.21831/jap.v5i2.32000>

- Utami, L., & Prasetyo, B. (2024). Dampak kelelahan kognitif akibat soal pseudo-HOTS terhadap penurunan performa tes matematika. *Jurnal Psikologi Pendidikan Matematika*, 7(1), 40–53. <https://doi.org/10.21009/jppm.v7i1.4100>
- Wahyuningsih, S., & Zulkarnain, Z. (2026). Re-kalibrasi instrumen asesmen daerah melalui analisis IRT multi-grup untuk menjamin keadilan evaluasi. *Jurnal Penelitian Evaluasi Pendidikan*, 28(1), 75–89. <https://doi.org/10.21831/jpep.v28i1.61000>
- Widiastuti, N., & Santoso, E. (2026). Menyeimbangkan kedalaman kognitif dan kerumitan kalkulasi dalam penyusunan soal HOTS matematika. *Jurnal Riset Pendidikan Matematika*, 13(1), 50–64. <https://doi.org/10.21831/jrpm.v13i1.64000>
- Yusuf, M., & Maryati, S. (2026). Pemetaan indikator 4C dalam instrumen pengujian matematika terstandar. *Jurnal Edukasi Matematika dan Sains*, 14(1), 15–28. <https://doi.org/10.21831/jems.v14i1.59000>