

Karakteristik Psikometrik Instrumen Tes Kemampuan Akademik (TKA) Matematika Materi Pecahan untuk Siswa Sekolah Dasar

Rosalia Annisa Putri¹, Tri Astari^{2*}

^{1,2} Universitas Jember, Jember, Indonesia

ARTICLE HISTORY

Received: 28 Juli 2026

Revised: 29 Agustus 2026

Accepted: 31 Agustus 2026

Keywords:

Discrimination Power,
Empirical Validity, Item
Analysis, Item Difficulty,
Reliability

Kata Kunci:

Daya Pembeda, Validitas
Empiris, Analisis Butir,
Tingkat Kesukaran,
Reliabilitas

DOI:

<https://doi.org/10.31949/ijie.v4i2.19827>

ABSTRACT

The quality of mathematics assessment depends not only on content coverage but also on how well each item functions empirically. This study aimed to analyze the psychometric characteristics of a fraction-based Mathematics Academic Competency Test (Tes Kemampuan Akademik/TKA) instrument for primary school students through empirical validity, reliability, item difficulty, and discrimination power. A quantitative descriptive item-analysis design was employed using responses from 27 sixth-grade students at SD Negeri Kaliboto Lor 07, Lumajang. The trial instrument consisted of 15 dichotomously scored items in multiple-choice and complex multiple-choice formats. Empirical validity was analyzed using Pearson item-total correlations with a critical value of $r = 0.381$ at $\alpha = .05$. Reliability was estimated using KR-20, while item difficulty and discrimination were analyzed within classical test theory. Ten of the 15 items met the empirical validity criterion, whereas five were eliminated. The KR-20 coefficient was 0.70. Among the retained items, seven were moderately difficult, two difficult, and one easy; four showed good discrimination and six showed fair discrimination. The findings indicate adequate preliminary psychometric quality, although broader validation with larger, multisite samples is still required before wider use.

ABSTRAK

Kualitas asesmen matematika tidak hanya ditentukan oleh keterwakilan materi, tetapi juga oleh fungsi empiris setiap butir. Penelitian ini bertujuan menganalisis karakteristik psikometrik instrumen Tes Kemampuan Akademik (TKA) Matematika materi pecahan untuk siswa sekolah dasar melalui validitas empiris, reliabilitas, tingkat kesukaran, dan daya pembeda. Penelitian menggunakan desain deskriptif kuantitatif berbasis analisis butir terhadap respons 27 siswa kelas VI SD Negeri Kaliboto Lor 07, Lumajang. Instrumen uji coba terdiri atas 15 butir berskor dikotomis dalam bentuk pilihan ganda dan pilihan ganda kompleks. Validitas empiris dianalisis menggunakan korelasi Pearson item-total dengan nilai kritis $r = 0,381$ pada $\alpha = 0,05$. Reliabilitas diestimasi menggunakan KR-20, sedangkan tingkat kesukaran dan daya pembeda dianalisis dalam kerangka teori tes klasik. Hasil menunjukkan 10 dari 15 butir memenuhi kriteria validitas empiris, sedangkan lima butir dieliminasi. Koefisien KR-20 sebesar 0,70. Dari 10 butir yang dipertahankan, tujuh berkategori sedang, dua sukar, dan satu mudah; empat memiliki daya pembeda baik dan enam cukup. Temuan menunjukkan kualitas psikometrik awal yang memadai, tetapi validasi pada sampel yang lebih besar dan lintas sekolah masih diperlukan sebelum instrumen digunakan secara lebih luas.

1. PENDAHULUAN

Tes Kemampuan Akademik (TKA) dirancang untuk menyediakan informasi capaian akademik murid secara terstandar. Pada jenjang SD/MI, TKA Matematika mengukur penguasaan fakta, konsep, prinsip, dan prosedur serta kemampuan menerapkan pengetahuan matematika dalam penyelesaian masalah. Kerangka TKA tidak hanya menekankan penguasaan konten, tetapi juga penggunaan representasi, penalaran, pemecahan masalah, dan koneksi matematis. Oleh karena itu, ketepatan interpretasi hasil TKA bergantung pada kualitas butir

yang digunakan untuk merepresentasikan kemampuan tersebut (Kementerian Pendidikan Dasar dan Menengah Republik Indonesia, 2025; Pusat Asesmen Pendidikan, 2025).

Pecahan merupakan salah satu materi penting dalam domain bilangan sekaligus menjadi sumber kesulitan bagi siswa sekolah dasar. Kesulitan dapat muncul ketika siswa menghubungkan berbagai representasi pecahan, melakukan operasi hitung, menafsirkan masalah kontekstual, maupun menentukan prosedur penyelesaian yang sesuai (Labibah et al., 2021; Meirina et al., 2024; Rohmah et al., 2024). Kesalahan penyelesaian masalah pecahan juga dapat terjadi pada tahap memahami informasi, membentuk model matematika, melakukan operasi, hingga menentukan jawaban akhir (Mustofa et al., 2022; Nailia et al., 2023; Wardhani et al., 2024). Kondisi tersebut menunjukkan bahwa butir TKA materi pecahan tidak cukup hanya sesuai dengan cakupan materi dan tuntutan kognitif, tetapi perlu menunjukkan karakteristik empiris yang memadai ketika direspons oleh siswa sasaran.

Dalam kerangka teori tes klasik, kualitas instrumen dapat ditelaah melalui beberapa bukti empiris yang saling melengkapi. Korelasi butir-total memberikan informasi mengenai hubungan skor suatu butir dengan skor tes, sedangkan reliabilitas menunjukkan konsistensi skor instrumen sebagai satu kesatuan. Tingkat kesukaran memberikan informasi mengenai proporsi peserta yang mampu menjawab suatu butir dengan benar, sementara daya pembeda menunjukkan kemampuan butir membedakan peserta dengan performa lebih tinggi dan lebih rendah (American Educational Research Association, 2014; DeCarlo, 2023; DeVellis, 2017). Analisis setelah uji coba diperlukan karena kesesuaian butir secara konseptual belum menjamin bahwa setiap butir akan berfungsi secara memadai pada kelompok sasaran (Boateng et al., 2018; Waty et al., 2024). Penggunaan validitas empiris, reliabilitas, tingkat kesukaran, dan daya pembeda secara bersama juga memberikan dasar yang lebih kuat dalam menilai kualitas dan menentukan tindak lanjut terhadap suatu butir (Jafar et al., 2024; Nurhasanah et al., 2024).

Sejumlah penelitian menunjukkan bahwa analisis karakteristik butir dapat digunakan untuk mengidentifikasi butir yang layak dipertahankan maupun yang memerlukan revisi atau eliminasi. Pendekatan tersebut telah diterapkan pada berbagai instrumen pendidikan, termasuk tes matematika dan tes hasil belajar berbentuk pilihan ganda (Ikhsanudin et al., 2023; Nisa, 2022; Ramatu & Kongnyuy, 2025). Kesukaran dan daya pembeda juga perlu ditafsirkan secara bersamaan karena hasil yang baik pada satu indikator tidak selalu diikuti oleh karakteristik yang sama pada indikator lainnya (Rezigalla et al., 2024). Kajian instrumen asesmen selanjutnya menekankan pentingnya dukungan bukti empiris yang memadai sebelum skor digunakan untuk menginterpretasikan kemampuan peserta didik (Coppi et al., 2023). Pada pendidikan dasar, analisis empiris setelah penelaahan awal instrumen juga dapat memperkuat dasar pengambilan keputusan mengenai penggunaan butir dalam pengukuran (Nurhasanah et al., 2024; Riyadi et al., 2024).

Berangkat dari kebutuhan tersebut, artikel ini memfokuskan kajian pada kualitas empiris instrumen TKA Matematika materi pecahan setelah tahap penelaahan awal instrumen. Hasil validasi ahli tidak dianalisis kembali dalam artikel ini agar pembahasan terarah pada fungsi butir berdasarkan respons siswa uji coba. Penelitian ini bertujuan menganalisis validitas empiris, reliabilitas KR-20, tingkat kesukaran, daya pembeda, serta keputusan akhir butir pada instrumen TKA Matematika materi pecahan untuk siswa sekolah dasar. Keterpaduan kelima bukti tersebut digunakan untuk menentukan butir yang dapat dipertahankan, direvisi, atau dieliminasi sebelum instrumen digunakan dalam penelitian utama. Dengan fokus tersebut, penelitian diharapkan memberikan bukti empiris awal mengenai kualitas instrumen sekaligus mendukung penyusunan asesmen matematika sekolah dasar yang lebih tepat dalam merepresentasikan kemampuan siswa.

2. METODE

Penelitian ini menggunakan desain deskriptif kuantitatif berbasis analisis karakteristik butir dalam kerangka teori tes klasik (*classical test theory*). Analisis difokuskan pada data uji coba instrumen Tes Kemampuan Akademik (TKA) Matematika materi pecahan dan tidak menggunakan data penelitian utama mengenai kesalahan siswa maupun gaya belajar. Uji coba dilaksanakan pada siswa di luar subjek penelitian utama sehingga respons yang dianalisis secara khusus digunakan untuk mengevaluasi kualitas instrumen sebelum digunakan dalam pengumpulan data utama.

Subjek uji coba terdiri atas 27 siswa kelas VI SD Negeri Kaliboto Lor 07, Lumajang. Instrumen yang diuji terdiri atas 15 butir, yaitu sembilan soal pilihan ganda dan enam soal pilihan ganda kompleks. Cakupan materi meliputi pecahan senilai, perbandingan dan pengurutan pecahan, relasi pecahan bias-desimal-persen, serta operasi pecahan. Setiap butir diberi skor dikotomis, yaitu 1 untuk jawaban benar dan 0 untuk jawaban salah. Karakteristik skor tersebut menjadi dasar penggunaan KR-20 untuk mengestimasi reliabilitas internal instrumen (Mufit & Syamsidar, 2022; Ningrum & Sudarwati, 2022).

Instrumen disusun berdasarkan cakupan materi pecahan dalam kerangka TKA dengan tuntutan kognitif C2 sampai C5. Kisi-kisi digunakan untuk memastikan keterwakilan materi, indikator kompetensi, tingkat kognitif, dan bentuk soal sebelum instrumen diuji secara empiris. Ringkasan kisi-kisi instrumen disajikan pada Tabel 1.

Tabel 1. Kisi-Kisi Instrumen TKA Matematika Materi Pecahan

No	Sub-elemen	Indikator soal	Level kognitif	Bentuk soal	No. butir
1	Pecahan Senilai	Disajikan gambar lingkaran yang diarsir, siswa dapat menentukan pecahan biasa yang senilai dengan gambar tersebut.	C2 (Menentukan)	PG	1
2	Pecahan Senilai	Disajikan gambar persegi panjang yang diarsir, siswa dapat menentukan pasangan pecahan senilai secara visual dan simbolik.	C3 (Menentukan)	PG	2
3	Pecahan Senilai	Disajikan masalah kontekstual berupa soal cerita tentang membagi kue, siswa dapat menganalisis pernyataan yang menunjukkan hubungan pecahan senilai.	C4 (Menganalisis)	PG Kompleks (Cerita)	10
4	Perbandingan dan Pengurutan	Disajikan empat pilihan pecahan biasa dengan penyebut berbeda, siswa dapat membandingkan nilai pecahan untuk menentukan yang paling besar.	C3 (Membandingkan)	PG	3
5	Perbandingan dan Pengurutan	Disajikan tiga pecahan biasa, siswa dapat mengurutkan tiga bilangan pecahan biasa dengan penyebut berbeda dari nilai terkecil ke terbesar.	C3 (Mengurutkan)	PG	4

No	Sub-elemen	Indikator soal	Level kognitif	Bentuk soal	No. butir
6	Perbandingan dan Pengurutan	Disajikan soal cerita terkait data volume minuman dalam bentuk pecahan, desimal, dan persen, siswa dapat menganalisis perbandingan yang benar.	C4 (Menganalisis)	PG Kompleks (Cerita)	11
7	Relasi Bentuk Pecahan	Disajikan pecahan biasa, siswa dapat mengubah bilangan pecahan biasa menjadi bentuk persen dengan tepat.	C3 (Mengubah)	PG	5
8	Relasi Bentuk Pecahan	Disajikan soal cerita terkait data isi botol dalam tiga bentuk pecahan berbeda, siswa dapat membuktikan kesamaan nilai.	C4 (Membuktikan)	PG Kompleks (Cerita)	12
9	Relasi Bentuk Pecahan	Disajikan soal cerita terkait berbagai macam bentuk pecahan, siswa dapat menganalisis besaran air dalam bentuk desimal, pecahan, dan persen untuk menentukan subjek yang minum paling banyak	C4 (Menganalisis)	PG	6
10	Relasi Bentuk Pecahan	Siswa dapat menentukan kesamaan nilai pada takaran berat benda untuk menentukan wadah yang kapasitasnya sesuai	C3 (Menentukan)	PG Kompleks (Cerita)	13
11	Operasi Pecahan	Siswa dapat menghitung hasil penjumlahan dua pecahan biasa yang memiliki penyebut sama.	C3 (Menghitung)	PG	7
12	Operasi Pecahan	Siswa dapat menghitung hasil perkalian dua pecahan biasa dan menyederhanakan hasilnya	C3 (Menghitung)	PG	8
13	Operasi Pecahan	Disajikan masalah pembagian pita, siswa dapat menganalisis prosedur perkalian dan pembagian pecahan dalam masalah kontekstual	C4 (Menganalisis)	PG Kompleks (Cerita)	14
14	Operasi Pecahan	Siswa dapat menyelesaikan masalah kontekstual penjumlahan pecahan biasa dengan penyebut berbeda terkait berat gula.	C4 (Menyelesaikan)	PG	9
15	Operasi Pecahan	Siswa dapat mengevaluasi langkah penyelesaian secara logis terkait persediaan awal dan penggunaan tepung.	C5 (Mengevaluasi)	PG Kompleks (Cerita)	15

Uji coba dilakukan dengan memberikan 15 butir instrumen kepada 27 siswa. Respons siswa selanjutnya diberi skor berdasarkan kunci jawaban dan dihipunkan menjadi data butir serta skor total. Data tersebut dianalisis secara bertahap melalui validitas empiris, reliabilitas, tingkat

kesukaran, dan daya pembeda. Urutan analisis tersebut digunakan agar keputusan terhadap kualitas instrumen mempertimbangkan bukti pada tingkat butir maupun konsistensi tes secara keseluruhan.

Validitas empiris setiap butir dianalisis menggunakan korelasi *Pearson Product Moment* antara skor butir dan skor total. Dengan jumlah peserta uji coba sebanyak 27 siswa, derajat kebebasan (*df*) sebesar 25 dan taraf signifikansi 5% digunakan nilai kritis (*r*) sebesar 0,381. Butir dinyatakan memenuhi kriteria validitas empiris apabila $r_{hitung} > r_{tabel}$. Sebaliknya, butir dengan $r_{hitung} \leq r_{tabel}$ dikategorikan belum memenuhi kriteria dan menjadi dasar untuk revisi atau eliminasi. Korelasi item-total digunakan sebagai bukti awal mengenai keterkaitan fungsi setiap butir dengan skor tes secara keseluruhan (Schober et al., 2018).

Setelah validitas empiris butir dianalisis, konsistensi internal instrumen diestimasi menggunakan formula Kuder-Richardson 20 (KR-20). Pemilihan KR-20 didasarkan pada karakter skor butir yang bersifat dikotomis. Koefisien reliabilitas diinterpretasikan berdasarkan kategori yang digunakan dalam penelitian sumber, yaitu 0,80-1,00 sangat reliabel, 0,60-0,80 reliabel, 0,40-0,60 cukup reliabel, 0,20-0,40 kurang reliabel, dan 0,00-0,20 tidak reliabel. Reliabilitas dipahami sebagai karakteristik konsistensi skor pada konteks sampel uji coba, sehingga interpretasinya tetap mempertimbangkan jumlah butir dan karakteristik peserta (Taber, 2018; Tavakol & Dennick, 2011). Kategori reliabilitas yang digunakan dalam penelitian sumber tercantum dalam dokumen metode.

Analisis berikutnya dilakukan terhadap tingkat kesukaran setiap butir. Indeks tingkat kesukaran ((*P*)) ditentukan berdasarkan proporsi siswa yang menjawab butir dengan benar terhadap seluruh peserta uji coba. Butir dikategorikan sukar apabila $P \leq 0,30$, sedang apabila $0,31 \leq P \leq 0,70$ dan mudah apabila $P \geq 0,70$. Analisis ini digunakan untuk mengetahui sebaran tingkat tuntutan butir sehingga instrumen tidak hanya terdiri atas soal dengan tingkat kesukaran yang sama.

Daya pembeda dianalisis untuk mengetahui kemampuan setiap butir dalam membedakan siswa dengan performa lebih tinggi dan lebih rendah. Indeks daya pembeda (*D*) dihitung berdasarkan perbedaan proporsi jawaban benar antara kelompok atas dan kelompok bawah. Nilai 0,71–1,00 dikategorikan sangat baik, 0,41-0,70 baik, 0,21-0,40 cukup, 0,00-0,20 kurang, sedangkan nilai negatif menunjukkan bahwa butir tidak memiliki fungsi pembeda yang memadai.

Keputusan akhir terhadap butir dilakukan secara integratif dengan mempertimbangkan hasil validitas empiris, reliabilitas instrumen, tingkat kesukaran, dan daya pembeda. Validitas empiris digunakan sebagai dasar utama untuk menentukan keterkaitan butir dengan skor tes, sedangkan tingkat kesukaran dan daya pembeda digunakan untuk menjelaskan karakteristik fungsi masing-masing butir. Reliabilitas digunakan untuk menilai konsistensi skor instrumen sebagai satu kesatuan. Dengan demikian, keputusan mempertahankan, merevisi, atau mengeliminasi butir tidak didasarkan pada satu indeks secara terpisah, melainkan pada keterpaduan bukti hasil analisis.

3. HASIL

Hasil analisis disajikan mengikuti urutan tujuan penelitian, yaitu validitas empiris, reliabilitas KR-20, tingkat kesukaran, daya pembeda, dan keputusan akhir butir. Urutan ini digunakan agar bukti pada tingkat butir dan konsistensi instrumen dapat dibaca secara sistematis. Analisis validitas empiris menunjukkan bahwa 10 dari 15 butir memiliki koefisien korelasi butir-total yang melampaui *r* kritis 0,381 pada taraf signifikansi 5%. Lima butir lainnya berada di bawah batas tersebut. Ringkasan hasil validitas empiris disajikan pada Tabel 2.

Tabel 2. Validitas Empiris Butir Instrumen TKA Matematika Materi Pecahan

Butir	r hitung	Keputusan
1	0,206	Eliminasi
2	0,528	Dipertahankan
3	0,238	Eliminasi
4	0,513	Dipertahankan
5	0,567	Dipertahankan
6	0,613	Dipertahankan
7	0,519	Dipertahankan
8	0,487	Dipertahankan
9	0,606	Dipertahankan
10	0,357	Eliminasi
11	0,384	Dipertahankan
12	0,575	Dipertahankan
13	0,424	Dipertahankan
14	0,283	Eliminasi
15	0,214	Eliminasi

Butir yang memenuhi kriteria validitas empiris dan dipertahankan adalah nomor 2, 4, 5, 6, 7, 8, 9, 11, 12, dan 13. Koefisien tertinggi terdapat pada Butir 6 ($r = 0,613$). Butir 11 memperoleh $r = 0,384$ sehingga tetap memenuhi kriteria, tetapi posisinya sangat dekat dengan r kritis. Sementara itu, Butir 1, 3, 10, 14, dan 15 tidak memenuhi kriteria; Butir 10 memiliki nilai yang paling dekat dengan batas keputusan, yaitu $r = 0,357$.

Setelah validitas empiris butir dianalisis, konsistensi internal instrumen dievaluasi menggunakan KR-20. Hasil perhitungan menghasilkan koefisien sebesar 0,70. Berdasarkan kriteria yang digunakan dalam penelitian sumber, nilai tersebut berada pada kategori reliabel. Ringkasannya disajikan pada Tabel 3.

Tabel 3. Reliabilitas Instrumen TKA Matematika Materi Pecahan

Instrumen	Koefisien KR-20	Kategori
TKA Matematika materi pecahan	0,70	Reliabel

Koefisien tersebut merupakan estimasi konsistensi internal skor yang diperoleh dari respons 27 siswa pada 15 butir berskor dikotomis. Hasil reliabilitas selanjutnya dibaca bersama karakteristik masing-masing butir, bukan sebagai dasar tunggal penetapan kualitas instrumen. Analisis tingkat kesukaran menunjukkan adanya variasi proporsi keberhasilan siswa pada 15 butir uji coba. Dua butir berada pada kategori mudah, sembilan butir pada kategori sedang, dan empat butir pada kategori sukar. Rincian indeks tingkat kesukaran disajikan pada Tabel 4.

Tabel 4. Tingkat Kesukaran Butir Instrumen

Butir	Indeks P	Kategori
1	0,85	Mudah
2	0,30	Sukar
3	0,70	Sedang
4	0,33	Sedang
5	0,70	Sedang
6	0,52	Sedang
7	0,41	Sedang
8	0,48	Sedang
9	0,26	Sukar
10	0,15	Sukar
11	0,78	Mudah
12	0,59	Sedang
13	0,37	Sedang
14	0,11	Sukar
15	0,41	Sedang

Setelah mempertimbangkan hasil validitas empiris, 10 butir yang dipertahankan terdiri atas satu butir mudah, tujuh butir sedang, dan dua butir sukar. Komposisi ini menunjukkan bahwa butir final tidak terkonsentrasi pada satu tingkat kesukaran saja.

Daya pembeda kemudian dianalisis untuk melihat kemampuan setiap butir dalam membedakan performa siswa pada kelompok atas dan kelompok bawah. Hasil analisis menunjukkan empat butir berkategori baik, tujuh butir berkategori cukup, dan empat butir berkategori kurang. Rincian daya pembeda disajikan pada Tabel 5.

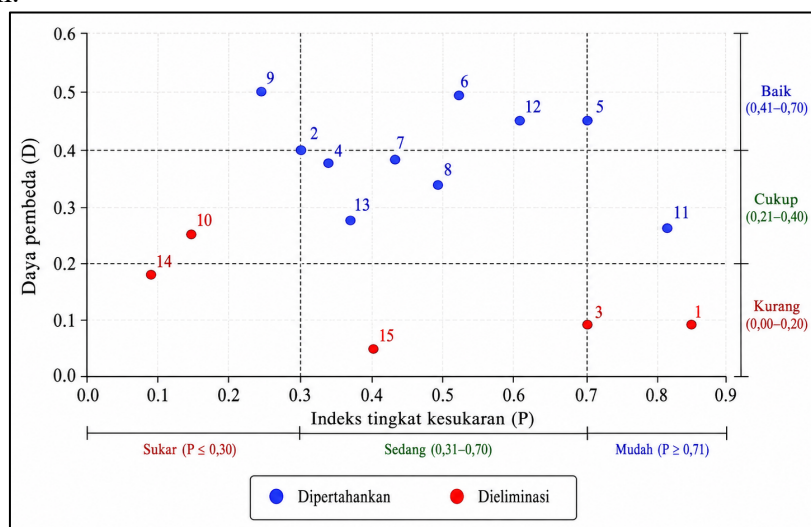
Tabel 5. Daya Pembeda Butir Instrumen

Butir	Indeks D	Kategori
1	0,09	Kurang
2	0,40	Cukup
3	0,09	Kurang
4	0,38	Cukup
5	0,45	Baik
6	0,49	Baik
7	0,38	Cukup
8	0,34	Cukup
9	0,50	Baik
10	0,25	Cukup
11	0,26	Cukup
12	0,45	Baik
13	0,28	Cukup
14	0,18	Kurang
15	0,05	Kurang

Pada 10 butir yang dipertahankan, empat butir memiliki daya pembeda kategori baik, yaitu Butir 5, 6, 9, dan 12, sedangkan enam butir lainnya berada pada kategori cukup. Tidak terdapat butir final dengan daya pembeda kurang atau negatif. Sementara itu, empat dari lima butir yang dieliminasi, yaitu Butir 1, 3, 14, dan 15, memiliki daya pembeda kurang. Butir 10 menunjukkan pola yang berbeda karena memiliki daya pembeda cukup ($D=0,25$), tetapi korelasi butir-totalnya sebesar 0,357 masih berada di bawah kritis 0,381.

Untuk memperjelas hubungan antara tingkat kesukaran dan daya pembeda pada setiap butir, hasil analisis divisualisasikan pada Gambar 2. Visualisasi tersebut menunjukkan posisi

relatif setiap butir berdasarkan kedua karakteristik sekaligus keterkaitannya dengan keputusan akhir instrumen.



Gambar 2. Distribusi Tingkat Kesukaran Dan Daya Pembeda Butir Instrumen

Gambar 2 menunjukkan variasi karakteristik antarbutir. Butir 1, 3, 14, dan 15 berada pada kategori daya pembeda kurang dan seluruhnya termasuk butir yang dieliminasi. Sebaliknya, butir yang dipertahankan memiliki daya pembeda cukup hingga baik dengan tingkat kesukaran yang bervariasi. Butir 10 menjadi kasus khusus karena daya pembedanya berada pada kategori cukup, tetapi tidak memenuhi kriteria validitas empiris. Pola tersebut menunjukkan bahwa keputusan terhadap suatu butir tidak cukup didasarkan pada satu indeks, melainkan perlu mempertimbangkan beberapa bukti empiris secara bersama.

Berdasarkan hasil analisis tersebut, keputusan akhir butir ditentukan dengan mempertimbangkan validitas empiris, tingkat kesukaran, dan daya pembeda. Sementara itu, koefisien KR-20 sebesar 0,70 digunakan untuk memberikan informasi mengenai konsistensi internal instrumen sebagai satu kesatuan. Hasil seleksi menghasilkan 10 butir yang dipertahankan dan lima butir yang dieliminasi, sebagaimana diringkas pada Tabel 6.

Tabel 6. Keputusan Akhir Butir Berdasarkan Bukti Empiris

Keputusan	Butir	Dasar Keputusan
Dipertahankan	2, 4, 5, 6, 7, 8, 9, 11, 12, 13	Memenuhi kriteria validitas empiris. Seluruh butir memiliki daya pembeda cukup atau baik dengan tingkat kesukaran yang bervariasi.
Dieliminasi	1, 3, 10, 14, 15	Tidak memenuhi kriteria validitas empiris. Butir 1, 3, 14 dan 15 juga memiliki daya pembeda kurang, sedangkan Butir 10 memiliki daya pembeda cukup tetapi korelasi butir totalnya berada di bawah batas 0,381

Berdasarkan Tabel 6, sebanyak 10 dari 15 butir dipertahankan sebagai instrumen final, sedangkan lima butir dieliminasi. Keputusan tersebut terutama didasarkan pada pemenuhan kriteria validitas empiris dan diperkuat oleh karakteristik tingkat kesukaran serta daya pembeda. Butir 10 memperlihatkan bahwa daya pembeda yang cukup tidak selalu diikuti oleh korelasi butir-total yang memenuhi kriteria. Instrumen final dengan 10 butir selanjutnya memiliki komposisi tingkat kesukaran yang terdiri atas satu butir mudah, tujuh butir sedang, dan dua butir sukar, dengan daya pembeda seluruhnya berada pada kategori cukup atau baik.

4. PEMBAHASAN

Temuan validitas empiris menunjukkan bahwa mayoritas butir memiliki keterkaitan yang memadai dengan skor total, tetapi terdapat variasi kekuatan hubungan antarbutir. Butir 11 memenuhi kriteria dengan $r = 0,384$, hanya sedikit di atas r kritis $0,381$, sedangkan Butir 10 memperoleh $r = 0,357$ dan tetap berada di bawah batas. Kondisi ini menunjukkan pentingnya konsistensi terhadap kriteria keputusan yang telah ditetapkan. Status suatu butir pada sampel uji coba juga perlu ditafsirkan secara hati-hati karena korelasi butir-total dapat berubah ketika instrumen digunakan pada kelompok dengan karakteristik berbeda (American Educational Research Association, 2014; DeCarlo, 2023). Butir yang berada dekat dengan batas keputusan lebih tepat dipantau atau diuji kembali daripada dinilai hanya berdasarkan selisih koefisien yang kecil (Rezigalla et al., 2024; Waty et al., 2024).

Reliabilitas KR-20 sebesar $0,70$ menunjukkan konsistensi internal yang berada pada kategori reliabel berdasarkan kriteria penelitian. KR-20 sesuai digunakan pada instrumen dengan skor dikotomis dan memberikan informasi mengenai konsistensi respons antarsekumpulan butir (Ningrum & Sudarwati, 2022; Rezigalla et al., 2024). Meskipun demikian, nilai tersebut belum tepat dipahami sebagai karakteristik yang tetap melekat pada instrumen. Reliabilitas bergantung pada konteks pengukuran, jumlah butir, serta keragaman peserta. Dengan 15 butir dan sampel uji coba sebanyak 27 siswa, koefisien $0,70$ lebih tepat diposisikan sebagai bukti konsistensi internal awal yang masih memerlukan konfirmasi pada sampel yang lebih besar dan beragam (Taber, 2018; Tavakol & Dennick, 2011).

Tingkat kesukaran menunjukkan bahwa instrumen uji coba memiliki variasi butir mudah, sedang, dan sukar. Pada 10 butir final, tujuh butir berada pada kategori sedang, satu mudah, dan dua sukar. Dominasi kategori sedang memberi ruang bagi instrumen untuk menangkap performa siswa di sekitar kemampuan menengah, sedangkan keberadaan butir mudah dan sukar membantu menjaga rentang tuntutan agar tidak terlalu sempit. Variasi tingkat kesukaran merupakan salah satu karakteristik penting dalam perangkat asesmen karena tes yang hanya didominasi satu kategori dapat memberikan informasi yang terbatas pada kelompok kemampuan tertentu (Ikhsanudin et al., 2023; Nisa, 2022; Ramatu & Kongnyuy, 2025).

Daya pembeda memberikan informasi tambahan yang tidak selalu sejalan dengan korelasi butir-total. Empat dari 10 butir final memiliki daya pembeda baik dan enam berkategori cukup. Butir 10 memperlihatkan pola yang penting: daya pembedanya masih berada pada kategori cukup ($D = 0,25$), tetapi korelasi butir-totalnya tidak memenuhi kriteria. Sebaliknya, Butir 11 memenuhi batas validitas empiris meskipun daya pembedanya hanya $0,26$. Perbedaan tersebut menunjukkan bahwa evaluasi butir tidak seharusnya bergantung pada satu indeks tunggal. Tingkat kesukaran, daya pembeda, dan korelasi butir-total perlu dibaca sebagai bukti yang saling melengkapi (DeCarlo, 2023; Rezigalla et al., 2024).

Keputusan akhir memperkuat pentingnya pembacaan integratif terhadap karakteristik butir. Empat dari lima butir yang dieliminasi, yaitu Butir 1, 3, 14, dan 15, tidak hanya gagal memenuhi kriteria validitas empiris, tetapi juga memiliki daya pembeda kurang. Butir 1 sangat mudah ($P = 0,85$) dengan daya pembeda $0,09$, sedangkan Butir 14 sangat sukar ($P = 0,11$) dengan daya pembeda $0,18$. Kedua contoh tersebut memperlihatkan bahwa tingkat kesukaran yang ekstrem tidak dengan sendirinya menghasilkan fungsi diskriminasi yang baik. Sementara itu, eliminasi Butir 10 menunjukkan bahwa daya pembeda cukup tidak dapat mengkompensasi korelasi butir-total yang belum memenuhi kriteria. Pola ini mendukung penggunaan beberapa sumber bukti secara bersama dalam menentukan butir yang dipertahankan, direvisi, atau dieliminasi (Nurhasanah et al., 2024; Rezigalla et al., 2024; Waty et al., 2024).

Sepuluh butir yang dipertahankan memberikan dasar empiris awal bagi penggunaan instrumen TKA Matematika materi pecahan dalam lingkup terbatas. Namun, temuan penelitian ini tetap berada dalam kerangka teori tes klasik dan berasal dari uji coba pada 27 siswa di satu

sekolah. Parameter korelasi butir-total, tingkat kesukaran, daya pembeda, dan reliabilitas dapat berubah ketika instrumen digunakan pada populasi berbeda (DeCarlo, 2023; Gorham, 2022). Pengujian lanjutan perlu melibatkan sampel lintas sekolah yang lebih besar, menelaah fungsi distraktor pada soal pilihan ganda, serta mempertimbangkan analisis Rasch atau *Item Response Theory* untuk memperoleh informasi yang lebih kuat mengenai kecocokan butir, tingkat kesukaran, reliabilitas, dan cakupan kemampuan (Darman et al., 2024; Nisfatulsanah & Sugiharto, 2024; Nurhasanah et al., 2025; Riyadi et al., 2024).

5. KESIMPULAN

Analisis validitas empiris menunjukkan bahwa 10 dari 15 butir instrumen TKA Matematika materi pecahan memenuhi kriteria korelasi butir-total, sedangkan lima butir tidak memenuhi batas yang ditetapkan. Reliabilitas KR-20 sebesar 0,70 berada pada kategori reliabel berdasarkan kriteria penelitian dan memberikan bukti konsistensi internal awal pada sampel uji coba. Dari 10 butir yang dipertahankan, satu butir berkategori mudah, tujuh sedang, dan dua sukar; empat butir memiliki daya pembeda baik dan enam berkategori cukup. Berdasarkan keterpaduan bukti tersebut, Butir 2, 4, 5, 6, 7, 8, 9, 11, 12, dan 13 dipertahankan, sedangkan Butir 1, 3, 10, 14, dan 15 dieliminasi.

Hasil ini mendukung penggunaan terbatas instrumen setelah seleksi butir, tetapi belum cukup untuk menyatakan karakteristik pengukurannya stabil pada populasi yang lebih luas. Jumlah peserta uji coba yang hanya 27 siswa dari satu sekolah serta posisi Butir 11 yang sangat dekat dengan batas validitas empiris menjadi pertimbangan penting. Pengujian berikutnya perlu melibatkan sampel yang lebih besar dan beragam, menelaah fungsi distraktor, serta menggunakan analisis Rasch atau *Item Response Theory* untuk memperkuat bukti kualitas dan ketepatan pengukuran.

DAFTAR PUSTAKA

- American Educational Research Association, A. P. A. & N. C. on M. in E. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Wardhani, D. A. P., Oktiningrum, W., Wibowo, A., Muslihasari, A., & Nurfianti, S. (2024). Analisis kesalahan siswa dalam menyelesaikan soal literasi numerasi berdasarkan Newman's Analysis Error serta upaya pemberian scaffolding pada siswa kelas V SD Negeri 1 Wonosari. *Primary Education Journals (Jurnal Ke-SD-An)*, 4(1), 114–122. <https://doi.org/10.36636/primed.v4i1.4175>
- Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quinonez, H. R., & Young, S. L. (2018). Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health*, 6. <https://doi.org/10.3389/fpubh.2018.00149>
- Coppi, M., Fialho, I., & Cid, M. (2023). Scientific literacy assessment instruments: A systematic literature review. *Educação Em Revista*, 39. <https://doi.org/10.1590/0102-4698237523-t>
- Darman, D. R., Suhandi, A., Kaniawati, I., Samsudin, A., & Wibowo, F. C. (2024). Development and Validation of Scientific Inquiry Literacy Instrument (SILI) Using Rasch Measurement Model. *Education Sciences*, 14(3), 322. <https://doi.org/10.3390/educsci14030322>
- DeCarlo, L. T. (2023). Classical Item Analysis from a Signal Detection Perspective. *Journal of Educational Measurement*, 60(3), 520–547. <https://doi.org/10.1111/jedm.12358>
- DeVellis, R. F. (2017). *Scale development: Theory and applications (4th ed.)*. Sage.
- Gorham, A. , & R. (2022). *Classical test theory*. Routledge.

- Ikhsanudin, I., Novaliah, N., Hidayatullah, H., & Almizi, M. (2023). A practical using of the QUEST program to analyze the characteristics of the test items in educational measurement. *JISAE: Journal of Indonesian Student Assessment and Evaluation*, 9(1), 37–43. <https://doi.org/10.21009/jisae.v9i1.31163>
- Jafar, L., Riyadi, & Ridwan, A. (2024). Penggunaan teori tes klasik untuk analisis butir soal pada asesmen. *IJSH: Indonesian Journal of Social and Humanities*, 2(3), 1–10. <https://jurnal.academiacenter.org/index.php/IJSH>
- Kementerian Pendidikan Dasar dan Menengah Republik Indonesia. (2025). *Peraturan Menteri Pendidikan Dasar dan Menengah Republik Indonesia Nomor 9 Tahun 2025 tentang Tes Kemampuan Akademik*. <https://Peraturan.Bpk.Go.Id/Details/321171/Permendikdasmen-No-9-Tahun-2025>.
- Labibah, N., Damayani, A. T., & Sary, R. M. (2021). Analisis Kesalahan Siswa Berdasarkan Teori Newman dalam Menyelesaikan Soal Cerita pada Materi Pecahan Kelas V Madrasah Ibtidaiyah. *Journal for Lesson and Learning Studies*, 4(2), 208–216. <https://doi.org/10.23887/jlls.v4i2.33265>
- Meirina, E. N., Riswari, L. A., & Kironoratri, L. (2024). Analisis kesalahan siswa kelas VI SD 3 Jekulo dalam menyelesaikan soal cerita materi pecahan. *JP2M (Jurnal Pendidikan Dan Pembelajaran Matematika)*, 10(1), 85–95. <https://doi.org/10.29100/jp2m.v10i1.5382>
- Mufit, F., & Syamsidar, S. (2022). Development of Four-Tier Multiple Choice Test Instrument to Identify Students' Concept Understanding of Newton's Law Material. *JIPF (Jurnal Ilmu Pendidikan Fisika)*, 7(2), 134. <https://doi.org/10.26737/jipf.v7i2.2369>
- Mustofa, R., Anif, S., & Muhibbin, M. (2022). Analisis Kesalahan dalam Menyelesaikan Soal Cerita Matematika pada Siswa Sekolah Dasar. *Jurnal Basicedu*, 6(3), 3492–3502. <https://doi.org/10.31004/basicedu.v6i3.2635>
- Nailia, V., Setiawan, D., & Purbasari, I. (2023). Studi Analisis Kesulitan Penyelesaian Soal Cerita pada Pembelajaran Matematika Sekolah Dasar. *JiIP - Jurnal Ilmiah Ilmu Pendidikan*, 6(4), 2595–2602. <https://doi.org/10.54371/jiip.v6i4.1878>
- Ningrum, A. S. B., & Sudarwati, E. (2022). Is the test sensible? Developing a critical reading test for Indonesian EFL learners. *JEELS (Journal of English Education and Linguistics Studies)*, 9(1), 209–227. <https://doi.org/10.30762/jeels.v9i1.4005>
- Nisa, C. (2022). Analisis Karakteristik Butir Soal Tes Ujian Kenaikan Kelas Pelajaran Matematika dengan Program Iteman. *YASIN*, 2(6), 862–870. <https://doi.org/10.58578/yasin.v2i6.1020>
- Nisfatulsanah, A., & Sugiharto, B. (2024). Comparing classical test theory and item response theory for respiratory system question instruments. *PIONIR: JURNAL PENDIDIKAN*, 13(3), 70. <https://doi.org/10.22373/pjp.v13i3.25435>
- Nurhasanah, Astari, T., Cahyanita, E., & Nuriman. (2025). Psychometric properties of a local wisdom-based science literacy instrument for Primary School Students: An Item Response Theory approach. *Jurnal Penelitian Pendidikan IPA*, 11(11), 965–975. <https://doi.org/10.29303/jppipa.v11i11.12973>
- Nurhasanah, Hidayatullah, Z., & Arif, Moh. B. S. (2024). Karakteristik instrumen tes literasi digital ditinjau dari validitas isi dan validitas empiris (kecocokan butir dengan model, reliabilitas, serta tingkat kesukaran butir). *Journal of Classroom Action Research*, 6(4), 916–923. <https://doi.org/10.29303/jcar.v6i4.9650>
- Pusat Asesmen Pendidikan. (2025). *Kerangka Asesmen Tes Kemampuan Akademik Jenjang SD/MI/Sederajat*. <https://pusmendik.kemendikdasmen.go.id/tka/tka/view/mata-pelajaran-wajib/sd/matematika>
- Ramatu, N., & Kongnyuy, P. (2025). An Analysis of Item Difficulty and Item Discrimination of Mathematics Multiple Choice Items (2023 and 2024) in Bamboutos Division of

- Cameroon. *East African Journal of Education Studies*, 8(3), 505–511. <https://doi.org/10.37284/eajes.8.3.3600>
- Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaifi, J., ALGhamdi, M. A., Al Ameer, A. Y., Yahia, A. I. O., Mohammed, O. A., & Adam, M. I. E. (2024). Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*, 24(1), 445. <https://doi.org/10.1186/s12909-024-05433-y>
- Riyadi, A., Susongko, P., & Munadi, M. (2024). Model asesmen literasi sains pada peserta didik Sekolah Dasar dengan aplikasi Model Rasch. *Journal of Education Research*, 5(3), 3044–3054. <https://doi.org/10.37985/jer.v5i3.1411>
- Rohmah, M., Hilyana, F. S., & Ermawati, D. (2024). Kemampuan Pemahaman Konsep Matematis Siswa Kelas V Materi Pecahan. *Al-Madrasah Jurnal Pendidikan Madrasah Ibtidaiyah*, 8(2), 708. <https://doi.org/10.35931/am.v8i2.3425>
- Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, 126(5), 1763–1768. <https://doi.org/10.1213/ANE.0000000000002864>
- Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, 48(6), 1273–1296. <https://doi.org/10.1007/s11165-016-9602-2>
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53–55. <https://doi.org/10.5116/ijme.4dfb.8dfd>
- Waty, E. R. K., Nengsih, Y. K., Handrianto, C., & Rahman, M. A. (2024). The quality of teacher-made summative tests for Islamic education subject teachers in Palembang, Indonesia. *Jurnal Cakrawala Pendidikan*, 43(1), 186–197. <https://doi.org/10.21831/cp.v43i1.53558>