



## Pengembangan Asesmen Kompetensi Minimum Literasi Numerik Bentuk Pilihan Ganda dengan Menerapkan *Item Response Theory* (IRT) untuk Peserta Didik Sekolah Dasar

Dyah Ayu Kumala Sari<sup>1</sup>, I Made Candiasa<sup>2</sup>, Ni Made Sri Mertasari<sup>3</sup>

<sup>1,2,3</sup> Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Pendidikan Ganesha  
[dyah@student.undiksha.ac.id](mailto:dyah@student.undiksha.ac.id), [candiasaimade@undiksha.ac.id](mailto:candiasaimade@undiksha.ac.id), [sri.mertasari@undiksha.ac.id](mailto:sri.mertasari@undiksha.ac.id)

### Abstrak

Penelitian ini bertujuan untuk mengembangkan instrumen Asesmen Kompetensi Minimum (AKM) literasi numerik berbentuk pilihan ganda dengan menerapkan *Item Response Theory* (IRT) model 2PL yang valid, reliabel, dan berkualitas untuk peserta didik sekolah dasar. Metode pengembangan mencakup tahap perancangan, validasi ahli, uji *one-to-one*, dan uji lapangan terhadap 100 peserta didik yang terbagi ke dalam tiga paket soal dengan masing-masing memuat 20 butir soal. Analisis berdasarkan teori tes klasik menunjukkan seluruh butir soal valid dengan koefisien korelasi poin-biserial melebihi  $r$ -tabel 0,195, serta reliabilitas KR-20 berkisar 0,79–0,81 (kategori tinggi). Sementara itu, estimasi parameter IRT 2PL membuktikan seluruh asumsi dasar terpenuhi dengan nilai RMSEA di bawah 0,08 pada ketiga paket. Parameter daya beda didominasi kategori baik hingga sangat baik, sedangkan tingkat kesukaran menunjukkan distribusi yang bervariasi dengan butir tertentu konsisten berkategori sukar. Uji validitas konkuren menggunakan PLS-SEM menunjukkan keselarasan signifikan antara skor instrumen dan hasil belajar matematika dengan koefisien jalur 0,592–0,615 ( $p=0,000$ ) serta  $R^2$  berkisar 0,350–0,378. Berdasarkan analisis keseluruhan, instrumen AKM literasi numerik berbasis IRT yang dikembangkan memiliki kualitas psikometris yang memadai dan layak digunakan sebagai alat ukur kemampuan literasi numerik peserta didik sekolah dasar.

**Kata Kunci:** Asesmen Kompetensi Minimum, Literasi Numerik, *Item Response Theory*, Pilihan Ganda, Sekolah Dasar.

### 1. Pendahuluan

Literasi numerik bukan sekadar kemampuan mekanis menghitung, melainkan kecakapan menggunakan konsep dan alat matematika untuk menyelesaikan masalah kontekstual dalam kehidupan sehari-hari (Hambleton et al., 1991). Urgensi penguatan kompetensi ini tercermin dari rendahnya mutu literasi matematika Indonesia di tingkat global. Berdasarkan laporan *Programme for International Student Assessment* (PISA) tahun 2022, Indonesia hanya memperoleh skor 366 (peringkat 70 dari 81 negara), jauh di bawah rata-rata OECD sebesar 472 poin. Menanggapi kesenjangan ini, pemerintah menerapkan Asesmen Kompetensi Minimum (AKM) sebagai pilar Asesmen Nasional untuk mengukur kemampuan literasi membaca dan numerik siswa (Pusat Asesmen dan Pembelajaran, 2021). Pada Level 3 AKM (fase kelas 5 dan 6 sekolah dasar), fokus numerasi ditekankan pada proses kognitif tinggi—memahami (*knowing*), menerapkan (*applying*), dan bernalar (*reasoning*)—melalui konten geometri (keliling/luas gabungan), bilangan, data, dan aljabar (pola bilangan dan rasio). Desain stimulusnya wajib mengintegrasikan tiga konteks utama, yaitu personal, sosial-budaya, dan saintifik (Pusat Asesmen dan Pembelajaran, 2021).

Meskipun format pilihan ganda (*multiple-choice*) lazim digunakan dalam evaluasi skala besar karena efisiensinya, efektivitas instrumen ini di lapangan masih terhambat secara psikometris. Banyak siswa sekolah dasar kesulitan memahami maksud tersirat dalam stimulus numerasi yang panjang. Di sisi lain, kemampuan fungsi distraktor (pengecoh) pada soal sering kali belum optimal dalam membedakan miskonsepsi berpikir siswa berkemampuan rendah tanpa membingungkan siswa berkemampuan tinggi. Permasalahan ini diperparah oleh keterbatasan guru dalam mengembangkan bank soal AKM Level 3 yang valid dan reliabel (Hanafi, 2022). Dari sisi analisis butir, evaluasi di sekolah dasar umumnya masih bertumpu pada Teori Tes Klasik (*Classical Test Theory* - CTT) yang sifatnya *sample-dependent* (Hambleton et al., 1991). Untuk mengatasi kelemahan tersebut, *Item Response Theory* (IRT) hadir sebagai pendekatan modern yang menghasilkan parameter butir objektif dan *sample-independent* (Hambleton et al., 1991).

Dalam pengujian IRT, model tiga parameter (3PL) kerap gagal konvergensi jika ukuran sampel terbatas, sementara model satu parameter (1PL/Rasch) terlalu restriktif karena menyamakan daya beda seluruh soal (Hambleton et al., 1991). Sebagai solusi, model dua parameter logistik (2PL) terbukti stabil dan akurat bahkan pada ukuran sampel kecil sekecil  $N = 100$  (Chalmers, 2012). Pengujian instrumen ini disempurnakan dengan analisis validitas konkuren melalui pendekatan *Partial Least Squares Structural Equation Modeling* (PLS-SEM) untuk melihat

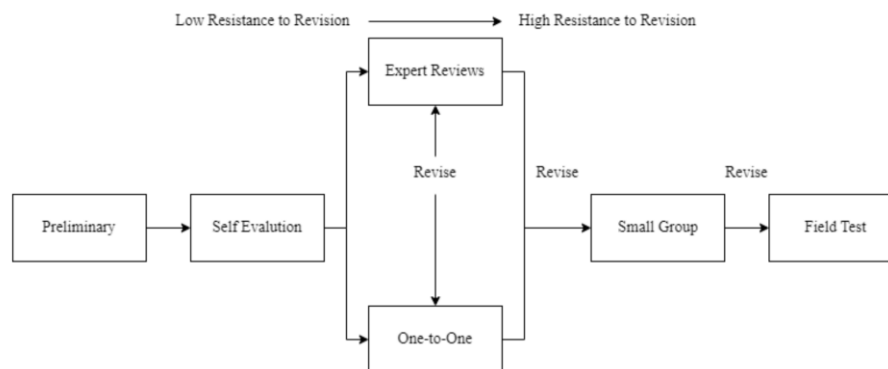
keselarasan instrumen dengan hasil belajar riil matematika siswa. Dengan memenuhi asumsi unidimensi, independensi lokal (Yen, 1984), dan invariansi parameter (Hambleton et al., 1991), probabilitas peserta  $i$  menjawab benar butir  $j$  dihitung berdasarkan model matematika berikut:

$$P_i(\theta_j) = \frac{1}{1 + e^{-Da_i(\theta_j - b_i)}} \quad (1)$$

Pada model di atas,  $a_j$  merepresentasikan parameter daya beda (*discrimination*) dari butir soal ke- $j$  dengan nilai ideal berkisar antara 0,0 hingga 2,0. Parameter  $b_j$  menunjukkan tingkat kesukaran (*difficulty*) butir ke- $j$  dalam satuan logit dengan kisaran ideal  $-2,0$  hingga  $+2,0$  (Hambleton et al., 1991). Adapun  $D$  merupakan konstanta skala bernilai 1,702 untuk menyelaraskan fungsi logistik agar mendekati fungsi ogive normal,  $e$  adalah bilangan eksponensial basis natural, dan  $\theta$  adalah parameter kemampuan siswa yang menyebar normal pada rentang nilai  $-3,0$  hingga  $+3,0$ . Melalui integrasi parameter  $a_j$  dan  $b_j$  yang ketat ini, penelitian ini bertujuan untuk mengembangkan dan menganalisis karakteristik parameter instrumen tes literasi numerik Level 3 sekolah dasar yang memenuhi prasyarat psikometris modern serta teruji validitas konkurensya terhadap hasil belajar matematika siswa.

## 2. Metode Penelitian

Penelitian ini menggunakan pendekatan penelitian dan pengembangan (*Research and Development - R&D*) dengan mengadaptasi model pengembangan instrumen *formative research* oleh Tessmer (1993). Prosedur pengembangan dilaksanakan melalui dua tahapan utama secara sistematis, yaitu tahap pendahuluan (*preliminary*) dan tahap evaluasi formatif (*formative evaluation*). Alur pengembangan instrumen tersebut divisualisasikan secara komprehensif pada Gambar 1.



Gambar 1. Diagram Alur Penelitian (Sumber: Tessmer, 1993)

Pada tahap *preliminary*, dilakukan studi literatur mengenai pemodelan IRT serta wawancara bersama guru matematika untuk memetakan kemampuan awal siswa. Tahap *self-evaluation* mencakup analisis kurikulum, karakteristik peserta didik kelas 5 sekolah dasar, dan hierarki konsep materi untuk menyusun cetak biru (*blueprint*) desain instrumen. Pada tahap ini, dihasilkan tiga paket soal paralel AKM literasi numerik Level 3 berbentuk pilihan ganda dengan 4 opsi jawaban (A, B, C, D) yang mencakup domain bilangan, aljabar, geometri-pengukuran, serta data-ketidakpastian.

Evaluasi formatif diawali melalui *expert review* oleh 7 pakar (2 dosen matematika Universitas Pendidikan Ganesha dan 5 guru sekolah dasar di Kabupaten Buleleng). Hasil penilaian pakar dianalisis menggunakan indeks validitas isi Aiken's  $V$  melalui Rumus 1.

$$V = \frac{\sum s}{n(c-1)} \quad (2)$$

dengan  $V$  adalah koefisien validitas isi Aiken,  $s$  adalah selisih antara nilai kategori yang dipilih pakar ke- $i$  ( $r_i$ ) dengan nilai kategori terendah ( $lo$ ),  $n$  adalah jumlah pakar penilai, dan  $c$  adalah angka kategori peringkat tertinggi yang dapat dipilih pakar. Butir soal dikategorikan valid dan memiliki derajat kelayakan tinggi jika nilai indeks  $V \geq 0,80$  (Soesana dkk., 2023). Butir yang memenuhi syarat dilanjutkan pada uji *one-to-one* kepada 9 peserta didik heterogen menggunakan pendekatan *cognitive walkthrough*, diikuti uji *small group* kepada 30–35 peserta didik di Kecamatan Buleleng menggunakan pendekatan *heuristics* untuk mengeliminasi ambiguitas kebahasaan.

Tahap akhir adalah *field test* (uji lapangan) kepada 100 peserta didik kelas 5 sekolah dasar di Kecamatan Buleleng untuk mengumpulkan data respons dikotomi (benar = 1, salah = 0). Analisis kualitas psikometris butir dikaji menggunakan dua pendekatan. Pendekatan pertama menggunakan Teori Tes Klasik (CTT) untuk menghitung korelasi poin-biserial ( $r_{pbis}$ ), reliabilitas Kuder-Richardson 20 (KR-20), indeks kesukaran klasik ( $p$ ) daya

pembeda klasik ( $ID$ ), dan indeks efektivitas fungsi pengecoh ( $IP$ ). Rumus-rumus perhitungan parameter CTT dipaparkan secara berurutan pada Rumus 2, Rumus 3, Rumus 4, Rumus 5, dan Rumus 6.

$$r_{pbis} = \frac{M_p - M_t}{SD_t} \sqrt{\frac{p}{q}} \quad (3)$$

$$r_{11} = \left(\frac{k}{k-1}\right) \left(1 - \frac{\sum pq}{SD_t^2}\right) \quad (4)$$

$$p = \frac{B}{N} \quad (5)$$

$$ID = \frac{B_A - B_B}{J} \quad (6)$$

$$IP = \frac{N_p}{\frac{N-B}{m-1}} \quad (7)$$

dengan  $r_{pbis}$  adalah koefisien korelasi poin-biserial butir,  $M_p$  adalah rata-rata skor total subjek yang menjawab benar pada butir tertentu,  $M_t$  adalah rata-rata skor total seluruh subjek,  $SD_t$  adalah standar deviasi skor total populasi sampel,  $p$  adalah proporsi subjek yang menjawab benar atau indeks kesukaran klasik,  $q$  adalah proporsi subjek yang menjawab salah ( $1 - p$ ),  $r_{11}$  adalah koefisien reliabilitas tes KR-20,  $k$  adalah kuantitas butir soal dalam instrumen,  $B$  adalah jumlah responden yang menjawab benar pada suatu butir,  $N$  adalah jumlah total responden,  $ID$  adalah indeks daya pembeda klasik,  $B_A$  adalah jumlah kelompok atas yang menjawab benar,  $B_B$  adalah jumlah kelompok bawah yang menjawab benar,  $J$  adalah jumlah sampel pada salah satu kelompok ekstrem (27% dari total sampel),  $IP$  adalah indeks pengecoh,  $N_p$  adalah jumlah peserta yang memilih opsi pengecoh tertentu, dan  $m$  adalah jumlah alternatif pilihan jawaban pada butir soal. Pengecoh dinyatakan berfungsi efektif jika nilai  $IP \geq 5\%$ . Kriteria klasifikasi parameter CTT disajikan secara ringkas pada Tabel 1.

Tabel 1. Klasifikasi Parameter CTT

Tingkat Kesukaran ( $p$ )		Daya Pembeda ( $ID$ )		Efektivitas Pengecoh per Soal	
Interval	Interpretasi	Interval	Interpretasi	Jumlah Berfungsi	Kriteria
$p < 0,30$	Sukar	$ID \geq 0,30$	Sangat Baik	3 Pengecoh	Sangat Baik
$0,30 \leq p \leq 0,70$	Sedang	$0,20 \leq ID \leq 0,29$	Baik	2 Pengecoh	Baik
$p > 0,70$	Mudah	$0,00 \leq ID \leq 0,19$	Sedang	1 Pengecoh	Kurang Baik
		$ID < 0,00$	Tidak Layak / Diganti	0 Pengecoh	Tidak Baik

Pendekatan kedua menerapkan *Item Response Theory* (IRT) model 2 Parameter Logistik (2PL) menggunakan bantuan perangkat lunak RStudio melalui *package* mirt dengan metode estimasi *Marginal Maximum Likelihood* (MML). Pengujian asumsi modern dilakukan secara ketat sebelum estimasi model, meliputi: (1) asumsi unidimensi menggunakan *Confirmatory Factor Analysis* (CFA) dengan syarat rasio *eigenvalue* komponen pertama dan kedua  $> 3:1$  atau varians faktor dominan tunggal mengekstrak informasi esensial instrumen; (2) asumsi independensi lokal melalui probabilitas gabungan lintas butir, yang membuktikan bahwa respons antar-item tidak saling berkorelasi pada tingkat kemampuan ( $\theta$ ) yang sama; dan (3) asumsi invariansi parameter melalui teknik korelasi linear parameter antar-subkelompok sampel. Fungsi probabilitas logistik IRT 2PL dihitung menggunakan Rumus 8.

$$P_i(\theta_j) = \frac{1}{1 + e^{-D a_i(\theta_j - b_i)}} \quad (8)$$

dengan  $P_i(\theta_j)$  adalah probabilitas peserta tes ke- $j$  dengan tingkat kemampuan  $\theta_j$  menjawab benar pada butir soal ke- $i$ ,  $e$  adalah konstanta bilangan eksponensial natural,  $D$  adalah faktor skala bernilai  $1,702$ ,  $a_i$  adalah parameter daya beda (*discrimination*) butir ke- $i$  dengan ambang batas ideal  $0 \leq a_i \leq 2$ , dan  $b_i$  adalah parameter tingkat kesukaran (*difficulty*) butir ke- $i$  dalam satuan logit dengan rentang kelayakan praktis  $-2,00 \leq b_i \leq +2,00$ .

Tahap akhir metode analisis menguji validitas konkuren instrumen terhadap hasil belajar matematika riil siswa (nilai ujian semester yang berjalan simultan) melalui pendekatan *Partial Least Squares Structural Equation Modeling* (PLS-SEM) menggunakan bantuan perangkat lunak SmartPLS. Pemodelan struktural (*inner model*) dijalankan secara terpisah pada tiga paket soal menggunakan indikator tunggal (*single indicator*), di mana skor total instrumen bertindak sebagai variabel eksogen dan nilai ujian bertindak sebagai variabel endogen. Karena sifat indikatornya yang tunggal, evaluasi *outer model* diabaikan (Hair dkk., 2019). Pengujian difokuskan pada kekuatan *path coefficient* ( $\beta$ ), nilai koefisien determinasi ( $R^2$ ), serta tingkat signifikansi hubungan prediktif melalui algoritma *bootstrapping* dengan ambang batas signifikansi *t-statistics*  $> 1,96$  pada *alpha* 5%.

### 3. Hasil dan Pembahasan

#### 3.1. Hasil Penelitian Pengembangan

Pengembangan instrumen evaluasi ini menghasilkan tiga paket soal paralel Asesmen Kompetensi Minimum (AKM) literasi numerik Level 3 untuk peserta didik kelas lima sekolah dasar. Setiap paket instrumen terdiri atas 20 butir soal pilihan ganda dengan empat opsi pilihan jawaban. Seluruh butir soal dirancang mengacu pada cetak biru *framework* AKM yang diterbitkan oleh Badan Penelitian dan Pengembangan dan Perbukuan Kementerian Pendidikan dan Kebudayaan, yang mengintegrasikan komponen konten, konteks, dan proses kognitif secara seimbang untuk mengukur kemampuan penalaran matematis siswa dalam situasi riil harian.

Pada tahapan pendahuluan, peneliti menghimpun berbagai referensi teoritis mengenai evaluasi literasi numerik berbasis *Item Response Theory* (IRT) serta menetapkan tiga level kognitif utama yang meliputi tahap mengetahui (*knowing*), menerapkan (*applying*), dan menalar (*reasoning*). Langkah ini diikuti dengan penentuan subjek dan lokasi penelitian pada empat satuan pendidikan di Kecamatan Buleleng, yaitu SD Negeri 1 Kaliuntu, SD Negeri 3 Kaliuntu, SD Negeri 4 Kaliuntu, dan SD Laboratorium Undiksha. Berdasarkan hasil wawancara mendalam dan observasi bersama guru kelas lima pada sekolah mitra, teridentifikasi kebutuhan mendesak terhadap instrumen evaluasi latihan yang bersifat kontekstual untuk melayih dan mengukur kemampuan penalaran siswa, mengingat perangkat penilaian yang tersedia di sekolah masih didominasi oleh Latihan soal prosedural standar.

Tahap evaluasi diri ditindaklanjuti dengan melakukan analisis kurikulum, analisis karakteristik perkembangan kognitif siswa kelas lima, dan analisis materi. Peneliti menyusun urutan hierarki konsep matematika yang mencakup empat ranah konten utama, yaitu bilangan, aljabar, pengukuran dan geometri, serta data dan ketidakpastian. Ranah-ranah tersebut kemudian dikembangkan ke dalam stimulus soal berbasis konteks personal, sosial-budaya, dan saintifik, yang dilengkapi dengan rubrik penskoran dikotomi.

Evaluasi formatif diawali dengan pelaksanaan *expert review* untuk menguji validitas isi produk. Proses ini melibatkan tujuh validator ahli yang terdiri atas tiga dosen jurusan matematika Universitas Pendidikan Ganesha, serta empat guru sekolah dasar berpengalaman di Kabupaten Buleleng. Para pakar memberikan penilaian kuantitatif pada lembar validasi yang mencakup aspek materi, konstruksi, dan bahasa menggunakan skala ordinal satu sampai empat. Indeks validitas isi dihitung menggunakan formula Aiken's V dengan hasil yang disajikan pada Tabel 2.

Tabel 2. Hasil Uji Validitas Isi Instrumen AKM Literasi Numerik Menggunakan Aiken

No. Butir	V Materi	V Konstruksi	V Bahasa	V Rata-rata	Kategori
1	1,0000	0,9048	1,0000	0,9683	Validitas Tinggi
2	1,0000	0,9048	0,9048	0,9365	Validitas Tinggi
3	1,0000	1,0000	0,9524	0,9841	Validitas Tinggi
4	1,0000	0,9524	0,9048	0,9524	Validitas Tinggi
5	1,0000	1,0000	0,9524	0,9841	Validitas Tinggi
6	1,0000	0,8571	0,8571	0,9048	Validitas Tinggi
7	1,0000	0,9524	1,0000	0,9841	Validitas Tinggi
8	1,0000	1,0000	1,0000	1,0000	Validitas Tinggi
9	1,0000	1,0000	1,0000	1,0000	Validitas Tinggi
10	1,0000	1,0000	1,0000	1,0000	Validitas Tinggi
11	1,0000	1,0000	1,0000	1,0000	Validitas Tinggi
12	1,0000	1,0000	0,9524	0,9841	Validitas Tinggi
13	0,9524	0,9524	0,8571	0,9206	Validitas Tinggi
14	1,0000	0,9524	1,0000	0,9841	Validitas Tinggi
15	1,0000	1,0000	1,0000	1,0000	Validitas Tinggi
16	1,0000	0,9524	1,0000	0,9841	Validitas Tinggi
17	1,0000	0,9524	1,0000	0,9841	Validitas Tinggi
18	1,0000	1,0000	1,0000	1,0000	Validitas Tinggi
19	0,9048	0,9524	1,0000	0,9524	Validitas Tinggi
20	1,0000	1,0000	1,0000	1,0000	Validitas Tinggi
Rata-rata	0,9929	0,9667	0,9690	0,9762	Validitas Tinggi

Berdasarkan data pada Tabel 2, pengujian kuantitatif menunjukkan bahwa seluruh butir soal pada ketiga paket instrumen memperoleh indeks validitas Aiken dengan rata-rata kumulatif sebesar 0,9762. Nilai koefisien pada ranah materi mencapai 0,9929, sedangkan ranah konstruksi sebesar 0,9667 dan ranah bahasa sebesar 0,9690. Nilai indeks parsial terendah ditemukan pada butir nomor 6 dengan skor 0,8571 untuk dimensi konstruksi dan bahasa, yang secara statistik tetap melampaui ambang batas minimum kelayakan validitas tinggi sebesar 0,80. Sementara itu, butir soal nomor 8, 9, 10, 11, 15, 18, dan 20 memperoleh koefisien sempurna sebesar 1,0000 dari seluruh pakar.

Peneliti juga melakukan perbaikan kualitatif berdasar catatan validator, meliputi penyederhanaan struktur kalimat pada butir nomor 2 dan 12, rekonstruksi redaksi total pada butir nomor 6, substitusi istilah teknis "freezer" menjadi "lemari pendingin" pada butir nomor 5 agar lebih akrab bagi siswa sekolah dasar, penghapusan pronomina "ibunya" menjadi "ibu" pada butir nomor 11, serta eliminasi penggunaan simbol variabel  $x$  pada butir nomor 13 untuk menghindari bias pemahaman pada tingkat operasional konkret anak usia sekolah dasar.

Setelah fase revisi produk awal selesai, peneliti melaksanakan uji *one-to-one* kepada sembilan peserta didik kelas lima SD Negeri 3 Kaliuntu yang dipilih secara heterogen berdasarkan rekomendasi guru wali kelas, yang merepresentasikan kelompok kemampuan akademik tinggi, sedang, dan rendah. Melalui pendekatan *cognitive walkthrough*, siswa diminta menyelesaikan instrumen secara mandiri sembari mengutarakan kendala keterbacaan teks dan interpretasi stimulus. Hasil uji perorangan mengonfirmasi bahwa petunjuk pengerjaan dan esensi stimulus dapat dipahami dengan baik oleh seluruh kelompok, meskipun kelompok berkemampuan rendah membutuhkan alokasi waktu komparatif yang lebih lama ketika mencerna data berbasis tabel atau grafik. Masukan dari fase ini digunakan untuk memperbesar resolusi visual dan dimensi spasial gambar stimulus.

Penyempurnaan visual tersebut dilanjutkan dengan uji coba kelompok kecil (*small group evaluation*) yang melibatkan 30 peserta didik kelas lima SD Negeri 4 Kaliuntu. Berdasarkan data respons siswa, diperoleh nilai rata-rata skor total instrumen sebesar 11,61 dari skor maksimum 20, dengan deviasi standar sebesar 3,92. Rentang sebaran skor yang lebar antara nilai minimum sebesar 4 dan nilai maksimum sebesar 18 memberikan indikasi awal bahwa perangkat tes memiliki kemampuan membedakan variasi kompetensi literasi siswa secara memadai, sehingga seluruh paket dinyatakan siap untuk diujikan pada skala luas.

### 3.2. Hasil Uji Coba Lapangan Menggunakan Teori Tes Klasik

Tahap uji coba lapangan (*field test*) dilaksanakan secara luring pada empat sekolah dasar di bawah naungan Kecamatan Buleleng dengan melibatkan 100 peserta didik kelas lima untuk masing-masing paket instrumen, dengan durasi pengerjaan dibatasi selama 50 menit. Data matriks respons dikotomi yang diperoleh dari total 300 peserta tes kemudian diproses untuk menguji konsistensi internal butir melalui teknik korelasi poin-biserial. Hasil kalkulasi menunjukkan bahwa nilai koefisien korelasi seluruh butir soal pada Paket 1, Paket 2, dan Paket 3 secara konsisten berada di atas nilai kritis koefisien korelasi tabel (*r-table*) pada tingkat signifikansi lima persen, sehingga seluruh butir soal secara empiris dinyatakan valid secara internal.

Pengujian reliabilitas perangkat tes dilakukan menggunakan indeks Kuder-Richardson 20 (KR-20) berbantuan program R. Estimasi konsistensi internal memberikan koefisien reliabilitas Paket 1 sebesar 0,79 dengan rentang interval kepercayaan 95% pada sebaran 0,72 hingga 0,84. Paket 2 menghasilkan koefisien reliabilitas sebesar 0,81 dengan interval kepercayaan berada pada area 0,75 hingga 0,86. Sementara itu, Paket 3 memperoleh nilai indeks reliabilitas sebesar 0,80 dengan cakupan interval kepercayaan dari 0,73 hingga 0,85. Mengacu pada kriteria standarisasi pengukuran, indeks yang diperoleh Paket 1 dan Paket 3 berada pada kuadran keandalan tinggi, sedangkan Paket 2 berada pada kuadran keandalan sangat tinggi, yang menegaskan stabilitas fungsional instrumen dalam mengevaluasi literasi numerik.

Karakteristik tingkat kesukaran klasik  $p$  dianalisis secara proporsional berdasarkan perbandingan jumlah peserta tes yang merespons benar terhadap ukuran sampel total. Hasil pemetaan indeks tingkat kesukaran untuk ketiga paket instrumen memiliki karakteristik pola yang identik. Paket 1 didominasi oleh 11 butir berkategori mudah, 7 butir berkategori sedang, dan 2 butir berkategori sukar. Paket 2 memuat 13 butir mudah, 5 butir sedang, dan 2 butir sukar, sedangkan Paket 3 memiliki 12 butir mudah, 6 butir sedang, dan 2 butir sukar. Butir nomor 2 dan nomor 15 secara konsisten terbukti memiliki tingkat kesulitan tinggi di seluruh paket dengan nilai  $p$  berkisar antara 0,25 hingga 0,30. Fenomena ini mengindikasikan bahwa sebagian besar butir soal dapat diakses dengan baik oleh siswa, namun tetap mempertahankan beberapa jangkar soal berkategori sukar untuk menguji batas atas penalaran numerik siswa.

Analisis indeks daya pembeda  $ID$  dijalankan dengan membandingkan proporsi jawaban benar antara 27% peserta tes kelompok atas dengan 27% kelompok bawah, sehingga ukuran sampel penentu untuk setiap subkelompok ekstrem berukuran 27 peserta didik. Hasil analisis daya pembeda menunjukkan kualitas diskriminasi butir yang memuaskan secara menyeluruh. Paket 1 memiliki 13 butir berkategori sangat baik, 4 butir berkategori baik, dan 3 butir berkategori sedang. Daya pembeda tertinggi pada Paket 1 ditemukan pada butir nomor 9 dan 12 dengan indeks koefisien sebesar 0,815. Paket 2 memiliki sebaran 13 butir sangat baik, 4 butir baik, dan 3 butir sedang, dengan titik puncak daya pembeda ekstrim dicapai oleh butir nomor 9 dengan nilai koefisien sebesar 0,963. Pada Paket 3, sebaran kualitas terdiri atas 14 butir sangat baik, 5 butir baik, dan 1 butir berkategori sedang. Karena tidak

DOI: <https://doi.org/10.69693/ijmst.v4i2.10443>

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

ditemukan butir dengan nilai daya pembeda bernilai negatif di seluruh paket, tidak ada butir soal yang dieliminasi dari sistem perangkat tes.

Fungsionalitas efektivitas pengecoh dievaluasi untuk mengukur daya tarik opsi jawaban salah dalam mengelabui peserta tes yang belum menguasai konsep. Batasan efektivitas pengecoh terpenuhi jika opsi tersebut dipilih oleh minimal lima persen dari total peserta tes. Distribusi efektivitas pengecoh untuk masing-masing paket tes ditemukan pola perilaku distraktor yang stabil dan konsisten di ketiga paket soal. Paket 1 menempatkan 14 butir pada kriteria sangat baik, 3 butir kriteria baik, dan 3 butir kriteria kurang baik. Pola ini berulang pada Paket 2 dengan 14 butir sangat baik, 2 butir baik, dan 3 butir kurang baik. Sementara itu, Paket 3 menunjukkan perbaikan kualitas dengan 15 butir berkategori sangat baik, 1 butir kategori baik, dan 3 butir kategori kurang baik. Butir nomor 3, 14, dan 20 secara konsisten terdeteksi berada pada kategori kurang baik di seluruh paket karena hanya memiliki satu opsi pengecoh yang berfungsi optimal. Opsi pengecoh pada ketiga nomor tersebut memerlukan rekonstruksi konten alternatif agar lebih homogen dan memiliki daya tarik bagi siswa yang menebak jawaban. Walau demikian, mayoritas distraktor pada seluruh paket soal telah bekerja secara efektif.

### 3.3. Hasil Estimasi Parameter Teori Respons Butir Modern (IRT 2PL)

Sebelum parameter karakteristik modern diestimasi menggunakan model IRT Dua Parameter Logistik (2PL), matriks respons diuji untuk memenuhi tiga asumsi dasar teori tes modern yang meliputi asumsi unidimensi, independensi lokal, dan invariansi parameter. Pengujian asumsi unidimensi diaplikasikan melalui teknik analisis faktor eksploratori dengan metode *minimum residual*, analisis faktor paralel, serta analisis visual grafis *scree plot* berbantuan program R. Kriteria pemenuhan unidimensi didasarkan pada besaran dominansi faktor tunggal primer di mana rasio nilai *eigenvalue* antara faktor pertama dan faktor kedua wajib melebihi proporsi tiga banding satu. Rangkuman empiris pengujian parameter dimensi dominan tunggal untuk ketiga paket instrumen disajikan pada Tabel 3.

Tabel 3. Hasil Uji Asumsi Unidimensional

Paket Soal	Nilai <i>SS Loadings</i> Faktor Utama	Proporsi Varians Terjelaskan (%)	Rentang Muatan Faktor Butir	Rasio <i>Eigenvalue</i> (Syarat > 3:1)	Status Asumsi
Paket 1	3,97	20%	0,11 s.d. 0,62	4 : 1	Terpenuhi
Paket 2	4,30	22%	0,16 s.d. 0,66	4,3 : 1	Terpenuhi
Paket 3	4,04	20%	0,09 s.d. 0,64	4 : 1	Terpenuhi

Berdasarkan data yang tertera pada Tabel 3 menunjukkan bahwa ketiga paket instrumen secara konsisten dan stabil hanya mengukur satu dimensi kemampuan laten utama, yaitu kemampuan literasi numerik siswa kelas lima sekolah dasar. Terpenuhinya syarat mutlak aspek unidimensi pada seluruh paket soal ini memperbolehkan peneliti untuk melanjutkan operasional analisis data ke tahap pengujian asumsi berikutnya.

Pengujian asumsi independensi lokal diaplikasikan untuk memastikan bahwa respons peserta didik terhadap suatu butir soal tidak dipengaruhi oleh respons mereka terhadap butir soal lainnya setelah mengontrol kemampuan laten ( $\theta$ ). Evaluasi ini dilakukan menggunakan statistik koefisien  $Q_3$  yang dihitung berdasarkan estimasi model IRT 2PL berbantuan paket mirt pada program R. Kriteria pemenuhan asumsi menetapkan bahwa nilai koefisien  $Q_3$  antar-pasangan butir harus berada di bawah ambang batas kritis sebesar 0,40 untuk menyatakan tidak adanya ketergantungan lokal yang berlebihan. Hasil ringkasan statistik pengujian independensi lokal untuk ketiga paket soal disajikan pada Tabel 4.

Tabel 4. Hasil Uji Asumsi Idependen Lokal

Paket Soal	Rentang Nilai $Q_3$	Nilai Median	Nilai Rata-Rata	Koefisien $Q_3$ Tertinggi (Pasangan Butir)	Status Asumsi
Paket 1	-0,390 s.d. 0,364	-0,031	-0,040	0,364 (Butir 19 dan 20)	Terpenuhi
Paket 2	-0,453 s.d. 0,289	-0,023	-0,041	0,289 (Butir 19 dan 20)	Terpenuhi
Paket 3	-0,453 s.d. 0,289	-0,037	-0,041	0,345 (Butir 19 dan 20)	Terpenuhi

Berdasarkan data ringkasan pada Tabel 4 dominasi nilai negatif pada indikator median dan rata-rata di seluruh paket mengonfirmasi bahwa secara umum tidak terdapat ketergantungan positif yang berlebihan antar-butir, sehingga asumsi independensi lokal dinyatakan terpenuhi untuk ketiga paket soal. Fenomena konsistensi nilai  $Q_3$  tertinggi pada pasangan butir nomor 19 dan nomor 20 di ketiga paket mengindikasikan adanya kedekatan karakteristik konteks stimulus pada bagian akhir tes, meskipun posisinya secara statistik masih sangat aman di bawah batas toleransi.

Asumsi terakhir yang wajib dipenuhi adalah uji invariansi parameter yang bertujuan menjamin bahwa parameter karakteristik butir yang tidak berubah meskipun diujikan pada kelompok sampel peserta didik yang berbeda. Sesuai dengan landasan teoritis teori tes modern, invariansi parameter tercapai apabila pola parameter tingkat kesukaran dan daya pembeda yang dihasilkan oleh kelompok subjek yang berbeda menunjukkan konsistensi yang tinggi tanpa adanya pergeseran kategori yang ekstrem. Dalam penelitian ini, pembuktian invariansi parameter dilakukan dengan membandingkan stabilitas sebaran parameter empiris dari ketiga kelompok peserta uji lapangan yang mengerjakan masing-masing paket soal paralel.

Hasil komparasi parameter tingkat kesukaran antarpaket membuktikan adanya konsistensi pola yang sangat kuat, di mana butir nomor 2 secara konsisten menempati kategori sukar dengan nilai indeks kesukaran klasik sebesar 0,25 pada Paket 1, sebesar 0,27 pada Paket 2, dan sebesar 0,27 pada Paket 3. Pola yang sama ditunjukkan oleh butir nomor 15 yang konsisten berkategori sukar dengan indeks kesukaran sebesar 0,29 pada Paket 1, sebesar 0,30 pada Paket 2, dan sebesar 0,29 pada Paket 3. Stabilitas ini juga ditemukan pada kelompok butir operasional lainnya, di mana butir nomor 4, 6, 7, 13, 14, dan 19 secara konsisten masuk dalam rentang kategori mudah di ketiga paket, sementara butir nomor 3, 5, 9, 10, dan 12 secara konsisten bertahan pada kategori sedang. Pada dimensi daya pembeda, tingkat konsistensi parameter juga terkonfirmasi lewat butir nomor 9 dan nomor 12 yang selalu memperoleh indeks diskriminasi tertinggi dengan kategori sangat baik di seluruh paket, sedangkan butir nomor 6 secara konsisten berada pada kategori sedang dengan nilai koefisien yang stabil pada angka 0,111. Tidak ditemukannya perubahan kategori parameter yang melompat secara ekstrem antar-kelompok sampel menegaskan bahwa estimasi karakteristik instrumen terbebas dari pengaruh variasi kemampuan subjek penilai. Dengan demikian, asumsi invariansi parameter resmi terpenuhi dan analisis dilanjutkan ke tahap inti estimasi parameter modern model 2PL.

### 3.4. Estimasi Parameter Karakteristik Butir Model IRT 2PL

Setelah seluruh asumsi dasar teori respons butir terpenuhi secara komprehensif, peneliti melakukan estimasi parameter karakteristik modern menggunakan model dua parameter logistik (2PL) berbantuan paket mirt pada program R. Pengepasan data dijalankan menggunakan metode estimasi kemungkinan maksimum marginal (*Marginal Maximum Likelihood Estimation*) yang relevan untuk menghasilkan parameter diskriminasi  $a$  dan parameter tingkat kesukaran  $b$  yang tidak bias melalui algoritma komputasi iteratif yang berulang hingga mencapai kondisi konvergen. Parameter diskriminasi  $a$  merepresentasikan daya pembeda butir dalam memisahkan siswa berkemampuan tinggi dan rendah, sedangkan parameter tingkat kesukaran  $b$  mengindikasikan posisi tingkat kemampuan ( $\theta$ ) yang dibutuhkan siswa untuk memiliki peluang sebesar 50 persen dalam menjawab butir soal dengan benar, dengan rentang nilai ideal berada pada batasan -2 hingga +2. Hasil kalkulasi parameter modern IRT 2PL untuk ketiga paket instrumen dirangkum secara ringkas pada Tabel 5.

Tabel 5. Hasil Estimasi Parameter IRT 2PL

Paket Soal	Rentang Parameter Daya Pembeda ( $a$ )	Rentang Parameter Kesukaran ( $b$ )
Paket 1	0,273 s.d. 3,309	-3,052 s.d. 0,998
Paket 2	0,397 s.d. 4,220	-2,669 s.d. 0,948
Paket 3	0,217 s.d. 3,173	-2,668 s.d. 0,969

Berdasarkan paparan data pada Tabel 8, hasil estimasi parameter pada Paket 1 mencatat nilai daya pembeda ( $a$ ) bergerak dari 0,273 hingga 3,309, dengan kemampuan diskriminasi tertinggi ditemukan pada butir nomor 13 sebesar 3,309 yang diikuti oleh butir nomor 7 sebesar 3,074 dan butir nomor 14 sebesar 3,000. Parameter tingkat kesukaran ( $b$ ) Paket 1 berada pada rentang -3,052 hingga 0,998, dengan dua butir terdeteksi berada di luar batas ideal karena terlalu mudah, yaitu butir nomor 6 dengan nilai  $b = -2,712$  dan butir nomor 20 dengan nilai  $b = -3,052$ . Untuk Paket 2, parameter daya pembeda ( $a$ ) berada pada rentang 0,397 hingga 4,220, di mana nilai tertinggi dicapai oleh butir nomor 19 sebesar 4,220, sedangkan parameter kesukaran ( $b$ ) bergerak dari -2,669 hingga 0,948 dengan butir nomor 6 dan nomor 8 teridentifikasi terlalu mudah karena memiliki nilai  $b$  di bawah -2. Pada Paket 3, nilai parameter  $a$  memiliki rentang dari 0,217 hingga 3,173 dengan daya pembeda tertinggi pada

butir nomor 19 sebesar 3,173, serta parameter  $b$  berkisar dari -2,668 hingga 0,969 dengan hanya butir nomor 6 yang konsisten berada di luar rentang ideal dengan koefisien kesukaran modern sebesar -2,668.

Secara akumulatif, pola estimasi parameter modern memperlihatkan konsistensi yang sejalan dengan analisis klasik, di mana butir nomor 6 secara konsisten, terbukti sebagai butir paling mudah di seluruh paket soal dengan nilai kesukaran modern yang ekstrem di bawah -2. Hal ini bersesuaian dengan indeks kesukaran klasik butir nomor 6 yang menyentuh angka 0,96. Kelompok butir dengan nilai diskriminasi tinggi seperti nomor 7, 13, dan 19 juga menunjukkan stabilitas lintas paket yang menegaskan fungsionalitasnya yang sangat kuat dalam mengklasifikasikan tingkat kompetensi penalaran matematis siswa secara presisi.

### 3.5. Analisis Kurva Karakteristik Butir (Item Characteristic Curve / ICC)

Kurva Karakteristik Butir (*Item Characteristic Curve / ICC*) merupakan representasi grafis fungsional yang menggambarkan hubungan tak-linear antara tingkat kemampuan laten peserta didik ( $\theta$ ) pada sumbu horizontal dengan probabilitas atau peluang menjawab benar pada sumbu vertikal berdasarkan parameter model IRT 2PL. Kurva karakteristik yang ideal ditunjukkan oleh bentuk fungsi sigmoid menyerupai huruf S dengan tingkat kemiringan yang tajam pada titik infleksi, yang menandakan sensitivitas tinggi butir soal dalam membedakan kapasitas siswa seiring dengan peningkatan skala kemampuannya.

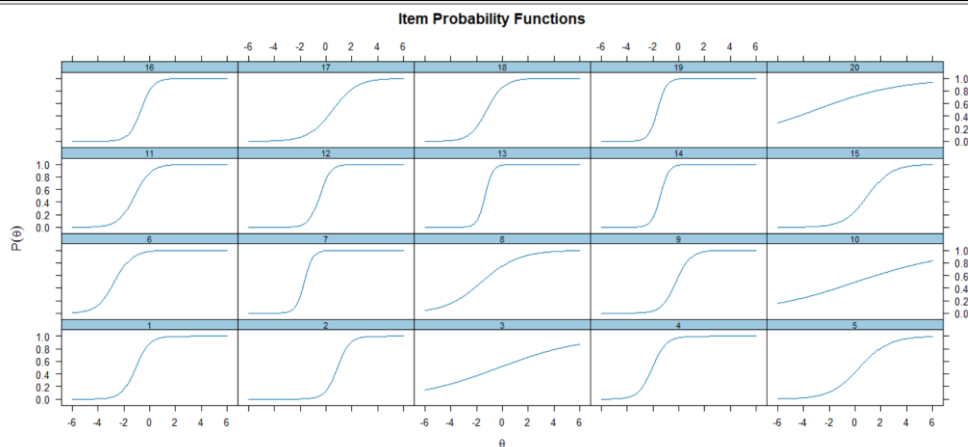
Untuk menjamin validitas dan kredibilitas hasil pengolahan data melalui algoritma paket mirt pada program R, peneliti melakukan verifikasi matematis secara manual menggunakan bantuan Microsoft Excel. Verifikasi difokuskan pada perhitungan nilai peluang menjawab benar  $P(\theta)$  dengan mensubstitusikan koefisien parameter  $a$  dan  $b$  hasil keluaran program ke dalam persamaan fungsi probabilitas logistik dua parameter. Formula dasar matematika yang digunakan untuk menghitung probabilitas menjawab benar pada tingkat kemampuan tertentu dengan menggunakan Rumus 8 dengan  $P(\theta)$  adalah probabilitas peserta didik menjawab benar pada tingkat kemampuan tertentu,  $D$  adalah konstanta penskalaan bernilai 1,702 atau 1,000,  $a$  adalah parameter daya pembeda butir hasil estimasi,  $b$  adalah parameter tingkat kesukaran butir hasil estimasi, dan  $\theta$  adalah tingkat kemampuan laten peserta didik.

Sebagai contoh visualisasi pembuktian keandalan komputasi, peneliti mengambil dua sampel butir ekstrim dari Paket 1 yang memiliki karakteristik bertolak belakang, yaitu butir nomor 1 dengan parameter daya pembeda tinggi ( $a = 1,887; b = -1,103$ ) dan butir nomor 20 dengan parameter daya pembeda rendah ( $a = 0,300; b = -3,052$ ). Perhitungan manual dilakukan pada titik kemampuan rata-rata populasi atau  $\theta = 0$  dengan menetapkan nilai konstanta penskalaan  $D = 1,000$ .

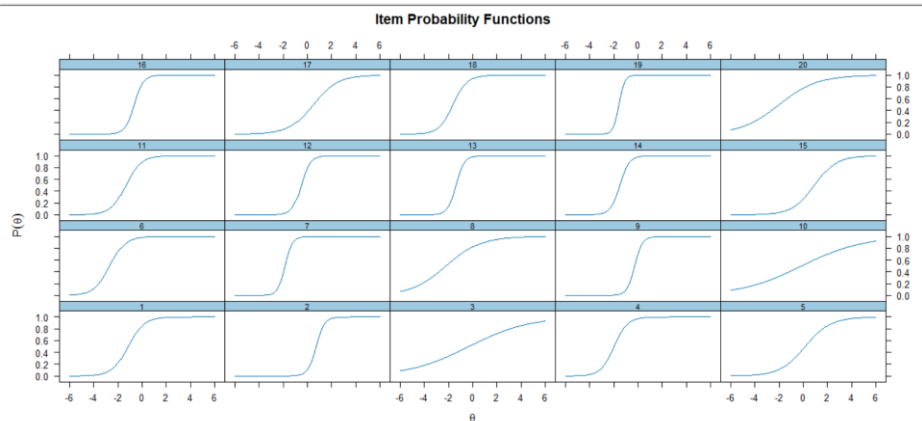
Langkah kalkulasi matematis manual untuk butir nomor 1 Paket 1 diawali dengan menghitung selisih antara tingkat kemampuan dan tingkat kesukaran sehingga diperoleh hasil  $(\theta - b) = 0 - (-1,103) = 1,103$ . Nilai selisih ini kemudian dikalikan dengan parameter diskriminasi butir untuk memproyeksikan nilai eksponen dasar sebesar  $a(\theta - b) = 1,887 \cdot 1,103 = 2,08136$ . Sesuai dengan struktur logistik, nilai eksponen diubah menjadi bertanda negatif untuk menghitung nilai fungsi eksponensial transenden yang menghasilkan nilai  $\exp(-2,08136) = 0,12476$ . Probabilitas akhir dihitung melalui pembagian satu dengan hasil penjumlahan satu ditambah nilai eksponensial tersebut, sehingga diperoleh nilai peluang final  $P(\theta = 0) = 1/(1 + 0,12476) = 0,889$ . Perhitungan ini mengindikasikan bahwa siswa dengan kemampuan rata-rata memiliki peluang sebesar 88,9% untuk menyelesaikan butir nomor 1 dengan benar.

Langkah operasional yang sama diterapkan untuk menguji butir nomor 20 Paket 1. Selisih kemampuan dan kesukaran menghasilkan nilai  $(\theta - b) = 0 - (-3,052) = 3,052$ , yang kemudian dikalikan dengan diskriminasi rendah sehingga menghasilkan nilai eksponen  $a(\theta - b) = 0,300 \cdot 3,052 = 0,9156$ . Konversi fungsi eksponensial dengan tanda negatif menghasilkan nilai  $\exp(-0,9156) = 0,40027$ . Melalui kalkulasi fungsi pembagian final, diperoleh peluang menjawab benar  $P(\theta = 0) = 1/(1 + 0,40027) = 0,714$ . Hasil ini menunjukkan bahwa butir nomor 20 memberikan peluang menjawab benar yang sudah tinggi (71,4%) bagi siswa berkemampuan rata-rata karena letak kesukaran butir yang berada sangat jauh di area negatif.

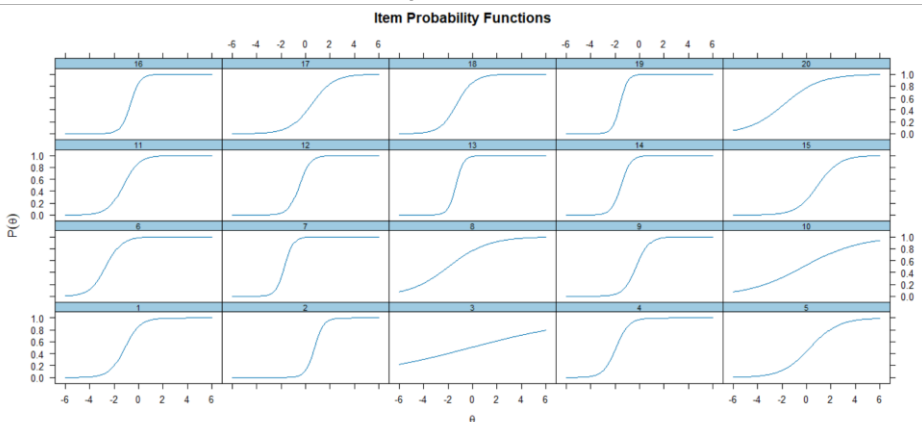
Seluruh baris data peluang hasil hitung manual untuk rentang  $\theta = -4$  hingga  $\theta = +4$  kemudian disusun dalam format matriks dan dibandingkan dengan tabel keluaran program R. Hasil pencocokan menunjukkan nilai probabilitas yang identik hingga empat angka di belakang koma, dan visualisasi bentuk kurva ICC yang diproduksi oleh Microsoft Excel terbukti sama persis dengan kurva grafis yang dihasilkan oleh program R. Validasi ini memberikan landasan kuat bagi peneliti untuk menampilkan profil ICC keseluruhan butir lintas paket yang merepresentasikan karakteristik fungsional perangkat tes.



Gambar 2. Kurva ICC Paket 1



Gambar 3. Kurva ICC Paket 2



Gambar 4. Kurva ICC Paket 3

Berdasarkan komparasi visual kurva ICC antarpaket, mayoritas butir soal pada Paket 1, Paket 2, dan Paket 3 berhasil membentuk pola kurva sigmoid S yang konsisten dan teratur. Hal ini menandakan adanya peningkatan probabilitas sukses merespons soal yang berbanding lurus dengan peningkatan kompetensi siswa. Butir-butir dengan diskriminasi tinggi seperti nomor 7, 13, 14, dan 19 di seluruh paket menampilkan kelandaian kurva yang sangat tegak di sekitar titik infleksinya, yang berarti perangkat tes memiliki sensitivitas ukur yang sangat responsif pada area kemampuan tersebut. Sebaliknya, butir nomor 3 dan nomor 10 menampilkan bentuk kurva yang cenderung mendatar, yang merefleksikan keterbatasan daya pembeda butir dalam mengelompokkan kapasitas siswa.

Ditinjau dari nilai rerata parameter diskriminasi kumulatif, Paket 2 menorehkan angka tertinggi sebesar 1,927 dengan kecenderungan visual kurva ICC yang paling tajam secara keseluruhan. Karakteristik ini membuat Paket 2 sangat ideal apabila fungsi instrumen diarahkan untuk tujuan seleksi atau pembedaan kemampuan dengan tingkat presisi yang tinggi. Di sisi lain, Paket 3 menunjukkan sebaran kurva yang paling seimbang antara kelandaian tinggi dan sedang dengan nilai rata-rata parameter  $a$  sebesar 1,798. Karakteristik sebaran Paket 3 ini menjadikannya sangat representatif dan adaptif untuk mengukur kemampuan siswa pada rentang populasi yang lebih luas dan

DOI: <https://doi.org/10.69693/ijmst.v4i2.10443>

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

inklusif, yang selaras dengan filosofi dasar pengembangan Asesmen Kompetensi Minimum sebagai alat pemetaan mutu pendidikan nasional.

### 3.6. Informasi Kecocokan Model (*Model Fit Information*)

Uji kecocokan model dilakukan untuk mengevaluasi derajat keselarasan atau proksimasi antara struktur teoritis model IRT 2PL yang diestimasi dengan matriks data empiris hasil uji lapangan. Evaluasi makro terhadap model secara keseluruhan ditentukan berdasarkan kombinasi indikator statistik yang meliputi nilai *Root Mean Square Error of Approximation* (RMSEA) dengan batas toleransi di bawah 0,08, nilai *Standardized Root Mean Square Residual* (SRMSR) dengan batas kelayakan di bawah 0,10, serta indeks *Tucker-Lewis Index* (TLI) dan *Comparative Fit Index* (CFI) dengan ambang batas ideal di atas 0,90. Pengujian tingkat mikro pada level butir dikonfirmasi melalui nilai signifikansi statistik  $S - X^2$ . Hasil pengujian kecocokan model makro untuk ketiga paket instrumen Paket 1 menghasilkan nilai  $M_2 = 250,088$  dengan nilai RMSEA sebesar 0,069 dan SRMSR sebesar 0,098, yang keduanya memenuhi kriteria batas kelayakan yang dapat diterima. Meskipun nilai TLI (0,877) dan CFI (0,890) berada sedikit di bawah ambang batas ideal 0,90, posisi angka tersebut sudah sangat dekat dengan batas optimal sehingga model tetap dinyatakan layak. Pada pengujian level butir Paket 1 melalui statistik  $S - X^2$ , sebagian besar butir soal terbukti cocok dengan nilai  $p > 0,05$ .

Untuk Paket 2, nilai RMSEA tercatat sebesar 0,078 dengan SRMSR tepat menyentuh angka 0,100, yang menunjukkan model masih berada dalam batas akseptabilitas empiris. Seluruh butir yang berhasil dihitung pada Paket 2 menunjukkan kecocokan yang sangat baik dengan nilai  $p - value$  melampaui taraf signifikansi 0,05. Sementara itu, Paket 3 menampilkan nilai RMSEA sebesar 0,073 dengan koefisien SRMSR sebesar 0,101 yang berada dalam toleransi wajar, di mana pada level individu butir tidak ditemukan adanya gejala *misfit* yang signifikan

Munculnya indikator nilai *Not a Number* (NaN) atau *Not Available* (NA) pada beberapa butir tertentu, seperti butir nomor 6 yang terjadi secara konsisten di ketiga paket soal serta butir nomor 7 dan nomor 19 pada Paket 2 dan Paket 3, disebabkan oleh keterbatasan variasi sebaran respons data empiris. Fenomena ini muncul akibat sangat tingginya proporsi peserta didik yang menjawab benar pada butir tersebut, seperti butir nomor 6 yang memiliki indeks kesukaran klasik sebesar 0,96, sehingga sistem komputasi kekurangan variasi data respons salah untuk mengkalkulasi matriks kovarians pada uji kesesuaian butir. Secara menyeluruh, pemenuhan indikator RMSEA di bawah 0,08 pada seluruh paket memberikan bukti kuat bahwa model IRT 2PL memiliki kecocokan yang memadai untuk digunakan sebagai basis penafsiran parameter instrumen evaluasi.

### 3.7. Uji Validitas Konkuren Instrumen Terhadap Hasil Belajar

Pengujian validitas konkuren dilaksanakan untuk mengevaluasi sejauh mana skor empiris yang dihasilkan oleh perangkat tes AKM literasi numerik yang dikembangkan memiliki tingkat keselarasan dan korelasi searah dengan indikator performa akademik siswa pada periode waktu yang bersamaan. Pengujian validitas eksternal ini dianalisis menggunakan metode *Partial Least Squares Structural Equation Modeling* (PLS-SEM) berbantuan perangkat lunak SmartPLS. Peneliti membangun tiga model jalur struktural independen untuk masing-masing paket soal, di mana skor total jawaban benar siswa diposisikan sebagai konstruk eksogen dengan indikator tunggal (*single indicator*), sedangkan nilai ulangan harian matematika siswa pada semester berjalan ditetapkan sebagai konstruk endogen. Karena model menggunakan indikator tunggal, evaluasi pengujian difokuskan pada analisis model struktural (*inner model*) yang mencakup koefisien jalur ( $\beta$ ), nilai koefisien determinasi ( $R^2$ ), serta tingkat signifikansi koefisien melalui prosedur pembentukan sampel ulang *bootstrapping*.

Hasil estimasi *bootstrapping* mengonfirmasi bahwa ketiga paket instrumen memiliki hubungan positif yang signifikan secara statistik dengan hasil belajar matematika siswa. Jalur struktural untuk Paket 1 menuju hasil belajar menghasilkan koefisien jalur sebesar 0,614 dengan nilai statistik  $T = 10,464$  dan nilai  $p = 0,000$ . Nilai koefisien determinasi  $R^2$  Paket 1 tercatat sebesar 0,378, yang mengindikasikan bahwa sebesar 37,8% variasi dari variabel hasil belajar matematika siswa dapat dijelaskan secara linier oleh skor perolehan tes Paket 1.

Untuk model struktural Paket 2, koefisien jalur menuju hasil belajar berada pada angka 0,615 dengan nilai statistik  $T = 9,760$  dan nilai  $p = 0,000$ , serta didukung oleh nilai  $R^2$  yang identik sebesar 0,378. Sementara itu, analisis jalur untuk Paket 3 menghasilkan koefisien arah sebesar 0,592 dengan nilai statistik  $T = 9,002$  dan nilai  $p = 0,000$ , dengan kontribusi nilai  $R^2$  sebesar 0,350 yang menunjukkan kemampuan prediktif sebesar 35,0% terhadap varians hasil belajar.

Kriteria signifikansi model struktural terpenuhi secara mutlak pada ketiga paket soal, ditandai oleh perolehan nilai statistik  $T$  yang melampaui nilai kritis tabel sebesar 1,96 serta nilai  $p - value$  yang konsisten berada di bawah taraf signifikansi lima persen ( $p < 0,05$ ). Berdasarkan standarisasi interpretasi structural, nilai koefisien determinasi yang berada pada rentang 0,350 hingga 0,378 masuk dalam kategori kekuatan asosiasi konkuren yang cukup karena telah melampaui batas minimum kriteria lemah sebesar 0,25. Sifat koefisien jalur yang bernilai positif dan signifikan memberikan bukti empiris bahwa semakin tinggi tingkat kemampuan literasi numerik yang ditunjukkan siswa melalui instrumen tes ini, semakin tinggi pula capaian hasil belajar matematika mereka. Hasil ini mengonfirmasi fungsionalitas instrumen yang dikembangkan dalam mengukur aspek kompetensi yang selaras dengan performa akademik riil siswa di sekolah.

### 3.8. Pembahasan Hasil Penelitian

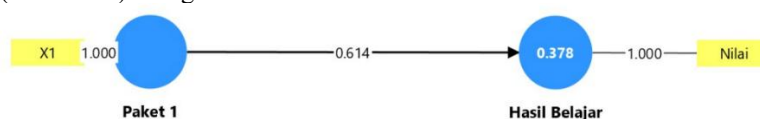
Penelitian pengembangan ini berfokus pada perancangan dan standarisasi instrumen evaluasi non-rutin berupa tiga paket soal paralel AKM literasi numerik Level 3 yang valid, andal, dan memiliki karakteristik butir yang baik. Standarisasi instrumen dilakukan melalui dua pendekatan komparatif, yaitu Teori Tes Klasik (*Classical Test Theory* - CTT) dan Teori Respons Butir Modern (*Item Response Theory* - IRT 2PL), guna menjamin akurasi dan objektivitas perangkat evaluasi sebelum digunakan secara luas untuk memetakan kemampuan penalaran matematis siswa sekolah dasar. Berdasarkan hasil uji validitas isi menggunakan indeks Aiken's V, keterbacaan instrumen dinilai sangat memadai dengan nilai rata-rata kumulatif mencapai 0,9762 (Mertasari, 2014). Hasil ini membuktikan bahwa adaptasi stimulus berbasis konteks personal, sosial-budaya, dan saintifik telah sesuai dengan struktur kurikulum dan tingkat perkembangan operasional konkret siswa kelas lima sekolah dasar. Masukan kualitatif dari para pakar dan umpan balik dari fase uji *one-to-one* terbukti efektif dalam meminimalkan bias keterbacaan bahasa, menyederhanakan kalimat stimulus yang terlalu kompleks, serta memperjelas tampilan grafis visual sebelum memasuki uji coba lapangan skala luas.

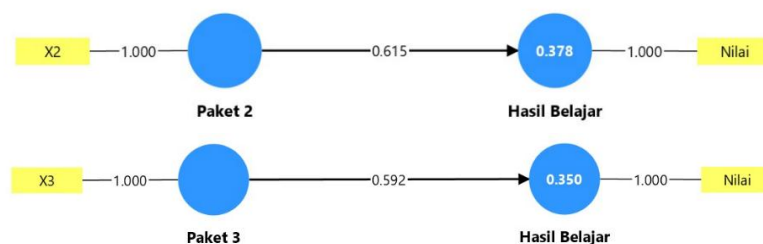
Kualitas instrumen berdasarkan analisis teori tes klasik menunjukkan tingkat validitas internal dan keandalan yang sangat memuaskan. Validitas internal dibuktikan oleh nilai koefisien korelasi poin-biserial seluruh butir yang sukses melampaui ambang batas kritis tabel  $r$  pada tingkat signifikansi lima persen. Tingkat konsistensi skor yang dihasilkan juga terbukti sangat stabil melalui perolehan koefisien reliabilitas KR-20 yang berkisar antara 0,79 hingga 0,81, yang menempatkan ketiga paket soal pada kuadran keandalan tinggi hingga sangat tinggi. Karakteristik ini memberikan jaminan bahwa instrumen mampu meminimalkan kesalahan pengukuran (*measurement error*) dan menghasilkan estimasi skor yang konsisten jika diujikan kembali pada kelompok siswa yang setara.

Ditinjau dari karakteristik tingkat kesukaran klasik ( $p$ ) dan daya pembeda ( $d$ ), ketiga paket instrumen menunjukkan keunggulan fungsional yang seimbang. Distribusi tingkat kesukaran didominasi oleh butir berkategori sedang hingga mudah, dengan mempertahankan butir nomor 2 dan nomor 15 sebagai jangkar soal berkategori sukar di seluruh paket. Struktur distribusi ini sangat ideal untuk menjaga motivasi pengerjaan siswa sembari tetap menyediakan instrumen penantang untuk memisahkan siswa pada batas atas kemampuan berpikir tingkat tinggi. Kualitas pemisahan ini diperkuat oleh analisis daya pembeda yang didominasi oleh kriteria baik hingga sangat baik, tanpa adanya butir soal yang memiliki koefisien diskriminasi negatif, sehingga tidak ada butir yang perlu dibuang dari sistem perangkat tes. Analisis efektivitas pengecoh menunjukkan perilaku distraktor yang stabil lintas paket, di mana sebagian besar opsi jawaban salah telah berfungsi secara efektif mengelabui siswa yang belum menguasai konsep dengan syarat keterpilihan minimal lima persen. Namun, butir nomor 3, 14, dan 20 secara konsisten diidentifikasi berada pada kategori kurang baik karena hanya memiliki satu opsi pengecoh yang berfungsi optimal, yang menjadi rekomendasi penting bagi perbaikan rekonstruksi opsi distraktor di masa depan agar lebih homogen.

Integrasi pendekatan teori tes modern model IRT 2PL memberikan konfirmasi karakteristik butir yang bebas dari sampel subjek (*sample-free*). Seluruh prasyarat dasar pengujian model modern yang meliputi asumsi unidimensi melalui analisis faktor paralel, asumsi independensi lokal melalui statistik koefisien  $Q_3$  Yen (1984), dan asumsi invariansi parameter terbukti terpenuhi secara meyakinkan pada ketiga paket soal. Estimasi parameter daya pembeda modern ( $a$ ) dan tingkat kesukaran modern ( $b$ ) menggunakan metode kemungkinan maksimum marginal menghasilkan parameter yang tidak bias, di mana butir nomor 13 dan nomor 19 secara konsisten dikonfirmasi memiliki daya pembeda tertinggi, sementara butir nomor 6 terbukti sebagai butir paling mudah dengan nilai parameter kesukaran modern yang berada di luar rentang teoretis ideal. Profil Kurva Karakteristik Butir (ICC) yang dihasilkan memperlihatkan bentuk fungsi sigmoid S yang sempurna pada mayoritas nomor soal, yang menandakan ketajaman akurasi pengukuran instrumen. Perbandingan kualitas antarpaket menunjukkan bahwa Paket 2 memiliki keunggulan pada aspek ketajaman diskriminasi individu sehingga sangat direkomendasikan untuk keperluan fungsi seleksi atau diagnosis presisi tinggi. Sementara itu, Paket 3 menunjukkan profil yang paling seimbang dan adaptif untuk mengukur kemampuan literasi pada rentang populasi yang luas, yang menjadikannya sebagai paket paling representatif dan inklusif untuk mendukung ketercapaian tujuan asesmen kompetensi minimum di tingkat sekolah dasar.

Kualitas eksternal dari instrumen yang dikembangkan akhirnya dibuktikan melalui uji validitas konkuren terhadap hasil belajar matematika peserta didik yang dianalisis menggunakan pendekatan *Partial Least Squares Structural Equation Modeling* (PLS-SEM) sebagai berikut.





Gambar 5. Hasil PLS *Algorithm* – Model Stuktural Paket 123

Path coefficients - Mean, STDEV, T values, p values					
	Original sample (O)	Sample mean (M)	Standard deviation (ST...)	T statistics ( O/STDEV )	P values
Paket 1 -> Hasil Belajar	0.614	0.613	0.059	10.464	0.000

Path coefficients - Mean, STDEV, T values, p values					
	Original sample (O)	Sample mean (M)	Standard deviation (ST...)	T statistics ( O/STDEV )	P values
Paket 2 -> Hasil Belajar	0.615	0.613	0.063	9.760	0.000

Path coefficients - Mean, STDEV, T values, p values					
	Original sample (O)	Sample mean (M)	Standard deviation (ST...)	T statistics ( O/STDEV )	P values
Paket 3 -> Hasil Belajar	0.592	0.590	0.066	9.002	0.000

Gambar 6. Hasil Bootstrapping – Path Coefficients Paket 123

Pengujian model jalur tunggal (*single path model*) menunjukkan bahwa ketiga paket instrumen memiliki koefisien jalur yang positif dan sangat kuat terhadap hasil belajar matematika siswa. Paket 1 mencatatkan koefisien sebesar 0,614, Paket 2 sebesar 0,615, dan Paket 3 sebesar 0,592. Seluruh nilai tersebut terbukti signifikan secara statistik pada *p-value* 0,000 dengan nilai *T*-statistik yang jauh melampaui ambang batas kritis 1,96. Hubungan ini dipertegas oleh nilai koefisien determinasi ( $R^2$ ) pada ketiga paket soal yang berkisar antara 0,350 hingga 0,378. Angka ini mengindikasikan bahwa variasi skor pada instrumen tes literasi numerik mampu menjelaskan sekitar 35% hingga 38% variasi dari hasil belajar matematika peserta didik pada momen yang sama.

Merujuk pada kriteria yang ditetapkan oleh Hair et al. (2019), kontribusi multivariat dalam rentang tersebut masuk ke dalam kategori cukup (*moderate*). Dalam konteks pengukuran psikologis dan evaluasi pendidikan, nilai kontribusi sebesar ini dinilai sangat memadai karena hasil belajar siswa di lapangan dipengaruhi oleh beragam faktor laten lain di luar kompetensi kognitif, seperti motivasi belajar dan kondisi lingkungan. Hal yang paling krusial dari temuan eksternal ini adalah konsistensi nilai koefisien jalur yang hampir setara di antara ketiga paket soal, yakni berkisar pada rentang sempit 0,592 hingga 0,615. Kesetaraan nilai prediktif ini menjadi bukti validasi silang (*cross-validation*) bahwa ketiga paket instrumen tidak hanya setara secara internal (berdasarkan karakteristik butir klasik dan teori tes modern), melainkan juga memiliki kualitas yang setara secara eksternal dalam memetakan kemampuan akademis siswa. Keberhasilan instrumen dalam menunjukkan validitas konkuren yang kokoh ini menegaskan bahwa perangkat tes yang dikembangkan memenuhi syarat standarisasi mutlak sebagai instrumen evaluasi yang berkualitas, valid, andal, dan memiliki keaslian (*authenticity*) tinggi untuk digunakan secara luas dalam memetakan literasi numerik siswa sekolah dasar.

#### 4. Kesimpulan dan Saran

Penelitian pengembangan ini berhasil menghasilkan instrumen tes literasi numerik Level 3 untuk siswa sekolah dasar yang terbagi ke dalam tiga paket soal paralel standar. Setiap paket memuat 20 butir soal pilihan ganda dengan empat alternatif jawaban yang disusun secara sistematis berdasarkan indikator literasi numerik melalui fase validasi ahli, uji *one-to-one*, dan uji lapangan. Karakteristik internal ketiga paket soal menunjukkan kualitas yang sangat memadai, di mana seluruh butir dinyatakan valid berdasarkan koefisien korelasi poin-biserial yang melampaui nilai *r*-tabel sebesar 0,195 pada taraf signifikansi lima persen. Tingkat konsistensi skor juga terbukti sangat tinggi dengan koefisien reliabilitas KR-20 yang terukur sebesar 0,79 untuk Paket 1, sebesar 0,81 untuk Paket 2, dan sebesar 0,80 untuk Paket 3.

Ditinjau dari karakteristik tingkat kesukaran klasik dan daya pembeda, ketiga paket instrumen menunjukkan sebaran tingkat kesulitan yang bervariasi dari sedang hingga mudah, dengan mempertahankan butir nomor 2 dan nomor 15 secara konsisten sebagai jangkar berkategori sukar. Seluruh butir soal terbukti memiliki fungsi pemisahan yang sehat tanpa adanya nilai daya pembeda negatif, di mana butir nomor 6 secara ajeg menunjukkan performa pembeda kategori sedang akibat indeks kesukarannya yang sangat tinggi mencapai 0,96 di seluruh paket. Integrasi pendekatan teori tes modern membuktikan terpenuhinya asumsi unidimensionalitas, independensi lokal, dan invariansi parameter pada model IRT 2PL secara meyakinkan. Estimasi parameter diskriminasi ( $a$ ) mengonfirmasi fungsionalitas butir nomor 7, 13, 14, dan 19 yang konsisten berdaya pembeda tinggi, sedangkan

parameter kesukaran ( $b$ ) mendeteksi butir nomor 6 terlalu mudah bagi peserta didik berkemampuan rata-rata. Kelayakan interpretasi parameter modern ini didukung secara empiris oleh pemenuhan nilai kecocokan model (RMSEA) di bawah batas toleransi 0,08 pada ketiga paket instrumen.

Kualitas eksternal dari perangkat evaluasi ini dikonfirmasi melalui pemenuhan validitas konkuren yang signifikan secara statistik terhadap hasil belajar matematika peserta didik. Analisis model jalur struktural PLS-SEM mencatat koefisien arah yang kuat yaitu sebesar 0,614 untuk Paket 1, sebesar 0,615 untuk Paket 2, dan sebesar 0,592 untuk Paket 3 dengan nilai kontribusi determinasi ( $R^2$ ) berkisar antara 35,0 persen hingga 37,8 persen. Kapasitas hubungan konkuren yang berada pada level cukup ini memberikan bukti nyata bahwa semakin tinggi tingkat penalaran numerik siswa pada instrumen ini, semakin baik pula performa akademik matematika mereka di kelas. Implikasi praktis dari penelitian ini adalah tersedianya paket evaluasi paralel standar yang siap diadopsi oleh guru dan pendidik sebagai referensi asesmen pemetaan literasi numerik di satuan pendidikan. Spekulasi penerapan instrumen ini dalam skala masif diharapkan dapat mendukung kesiapan sekolah dasar dalam menghadapi Asesmen Nasional secara lebih terstruktur dan berbasis data riil karakteristik butir.

Berdasarkan temuan selama penelitian, beberapa saran diajukan sebagai tindak lanjut perbaikan instrumen di masa depan. Bagi peneliti selanjutnya yang ingin mengembangkan instrumen sejenis, disarankan untuk melakukan rekonstruksi operasional terhadap butir nomor 6 agar tingkat kesukarannya lebih proporsional serta menguji validitas konkuren menggunakan kriteria eksternal yang lebih selaras seperti nilai ujian akhir semester yang diambil dalam waktu yang benar-benar bersamaan. Pengambilan sampel berikutnya juga perlu mempertimbangkan homogenitas asal sekolah peserta uji untuk menjamin kesetaraan skala nilai yang dibandingkan. Terakhir, disarankan bagi pengembangan lanjutan untuk melibatkan jumlah sampel yang jauh lebih besar dan heterogen agar diperoleh estimasi parameter IRT yang lebih stabil untuk generalisasi yang lebih luas, serta mempertimbangkan penggunaan model tiga parameter logistik (3PL) guna mengontrol bias dari faktor menebak (*guessing*) peserta didik.

## Reference

- Andikayana, D. M. (2021). *Pengembangan Instrumen Asesmen Kompetensi Minimum (Akm) Literasi Membaca Level 2 Untuk Siswa Kelas 4 Sd Tesis*. [Universitas Pendidikan Ganesha]. (Tesis Tidak Diterbitkan)
- Apriani, P. N. (2022). *Perbandingan Efektivitas Model Teori Respon Butir Satu*. Universitas Pendidikan Ganesha (Skripsi Tidak Diterbitkan).
- Arifin, Z., & Janan, T. (2023). Analisis Kesalahan Siswa Dalam Pemecahan Soal Asesmen Kompetensi Minimum Numerasi Berdasarkan Teori Polya. *Al Jabar: Jurnal Pendidikan Dan Pembelajaran Matematika, Volume 2 Nomor 2*.
- Augustia, A. D., Agustia, C. N., Azzahra, D., Hakim, L., & Pratiwi, V. (2025). Analisis Validitas Dan Reliabilitas Soal Pilihan Ganda Dengan Menggunakan Software Anates Pada Mata Pelajaran Perpajakan. In *Akademik: Jurnal Mahasiswa Ekonomi & Bisnis* (Vol. 5, Issue 1).
- Azid, A., Zamnah, L. N., & Solihah, D. S. (2023). *Mengapa Literasi Matematis Penting Dan Diperhatikan?* 3(1).
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates. [https://api.pageplace.de/preview/DT0400.9781135681470\\_A23807142/preview-9781135681470\\_A23807142.pdf](https://api.pageplace.de/preview/DT0400.9781135681470_A23807142/preview-9781135681470_A23807142.pdf)
- Fani, M. N. A. (2022). *Pengembangan Instrumen Evaluasi Tipe Tes Pilihan Ganda Berbasis Hots Mata Pelajaran Pai Dan Budi Pekerti Kelas 5 Sekolah Dasar Skripsi*. Universitas Islam Negeri Prof. K.H. Saifuddin Zuhri Purwokerto.
- Fernanda, J. W., Luthifiana, V., & Khoiril Akhyar, M. (2022). Analisis Partial Least Square Structural Equation Model (Pls-Sem) Untuk Pemodelan Penerimaan Sistem Jaringan Informasi Bersama Antar Sekolah (Jibas). *J Statistika, 15*(2), 292–297. [www.unipasby.ac.id](http://www.unipasby.ac.id)
- Firdausa, W. F., Denawanti, D., & Yulianti. (2025). Analisis Butir Soal Pilihan Ganda Penilaian Akhir Semester Pada Pembelajaran Tema 8 Praja Muda Karana Kelas Iii. *Jurnal Sadewa: Publikasi Ilmu Pendidikan, Pembelajaran Dan Ilmu Sosial, 3*(2), 129–151. <https://doi.org/10.61132/Sadewa.V3i2.1727>
- Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When To Use And How To Report The Results Of Pls-Sem. *European Business Review, 31*(1), 2–24. <https://doi.org/10.1108/Ebr-11-2018-0203>
- Hair, J., Hult, T., Ringle, C., Sarstedt, M., Danks, N., & Ray, S. (2021). *Partial Least Squares Structural Equation Modeling (Pls-Sem) Using R*. Classroom Companion: Business. <https://doi.org/10.1007/978-3-030-80519-7>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory Principles And Application*. Kluwer Nijhoff Publishing.
- Hanafi, H. (2022). Kendala guru dalam menyusun instrumen penilaian berbasis asesmen kompetensi minimum (AKM) di sekolah dasar. *Jurnal Elemen, 8*(1), 45–58.
- Husna, N. M., Isnarto, I., Suyitno, A., & Shodiqin, A. (2022). Integrasi Literasi Numerasi Dalam Pembelajaran Matematika Di Sekolah. *Prosiding Seminar Nasional Pascasarjana, 841–845*.
- Kemendikbud. (2020). *Akm Dan Implikasinya Pada Pembelajaran*. [https://pusmendik.kemdikbud.go.id/An/Page/Asesmen\\_Kompetensi\\_Minimum](https://pusmendik.kemdikbud.go.id/An/Page/Asesmen_Kompetensi_Minimum)
- Kementerian Pendidikan dan Kebudayaan. (2021). *Framework Asesmen Kompetensi Minimum (AKM)*. Badan Penelitian dan Pengembangan dan Perbukuan.
- Khakima, L. N., Fatimah, S., Zahra, A., Marlina, L., Abdullah, Z., & Pekalongan, I. (2022). *Prosiding Semai Seminar Nasional Pgm 2021 Penerapan Literasi Numerasi Dalam Pembelajaran Siswa Mi/Sd*. <http://proceeding.lainpekalongan.ac.id/Index.Php/Semai-775->
- Kopp, J., & Spoden, C. (2020). An optimized Bayesian hierarchical two-parameter logistic model for small-sample item calibration. <https://pubmed.ncbi.nlm.nih.gov/32536732/>
- Magdalena, I., Fauziah, S. N., Faziha, S. N., & Nupus, F. S. (2021). Analisis Validitas, Reliabilitas, Tingkat Kesulitan Dan Daya Beda Butir Soal Ujian Akhir Semester Tema 7 Kelas Iii Sdn Karet 1 Sepatan. In *Bintang: Jurnal Pendidikan Dan Sains* (Vol. 3, Issue 2). <https://ejournal.stitpn.ac.id/Index.Php/Bintang>
- Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement And Evaluation In Education And Psychology*. Ted Buchholz.
- Mertasari, N. M. S., & Candiasa, I. M. (2022). Formative Evaluation Of Digital Learning Materials. *Journal Of Education Technology, 6*(3), 507–514. <https://doi.org/10.23887/Jet.V6i3>
- Mertasari, N. M. S. (2021). *Pengujian Instrumen Penelitian Kuantitatif*. Rajawali Pers.
- Nadiem Makarim. (2019). *Merdeka Belajar Episode 1* [Broadcast]. Kementerian Pendidikan Dasar Dan Menengah Republik Indones.
- Naga, Dali S. (1992). *Pengantar Teori Sekor pada Pengukuran Pendidikan*. Gunadarma.

DOI: <https://doi.org/10.69693/ijmst.v4i2.10443>

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

- Pisa. (2023). *Pisa 2022 Results Factsheets Indonesia Pube*. <https://Oecdch.Art/A40de1dbaf/C108>.
- Radja, S. P., Bano, V. O., & Ina, A. T. (2023). Analisis Kualitas Butir Soal Hasil Belajar Peserta Didik Berdasarkan Tingkat Kesukaran, Daya Beda Dan Efektivitas Pengecoh Di Sman 1 Pandawai. *Jurnal Edusavana, Vol.1, No.1*.
- Rahadi, D. R. (2023). *Pengantar Partial Least Squares Structural Equation Model (Pls-Sem)*. Penerbit Lentera Ilmu Madani.
- Rahman, S. (2021). *Pascasarjana Universitas Negeri Gorontalo Prosiding Seminar Nasional Pendidikan Dasar "Merdeka Belajar Dalam Menyambut Era Masyarakat 5.0" Pentingnya Motivasi Belajar Dalam Meningkatkan Hasil Belajar*.
- Ringle, C. M., Wende, S., & Becker, J.-M. (2024). *Smartpls 4*. Smartpls.
- Rohim, D. C. (2021). Konsep Asesmen Kompetensi Minimum Untuk Meningkatkan Kemampuan Literasi Numerasi Siswa Sekolah Dasar. *Jurnal Varidika, 33(1)*, 54–62. <https://doi.org/10.23917/Varidika.V33i1.14993>
- Rohmad. (2017). *Pengembangan Instrumen Evaluasi Dan Penelitian* (1st Ed.). Kalimedia.
- Sa'dia, H. (2021). *Analisis Kemampuan Literasi Numerasi Ditinjau Dari Pengetahuan Metakognisi Siswa Dalam Menyelesaikan Soal Pisa Konten Space And Shape* [Universitas Islam Negeri Sunan Ampel Surabaya]. [http://digilib.uinsa.ac.id/49273/2/Holifatul%20sa%27dia\\_D74217045.Pdf](http://digilib.uinsa.ac.id/49273/2/Holifatul%20sa%27dia_D74217045.Pdf)
- Sugiyono. (2019). *Metode Penelitian Kualitatif, Kuantitatif, Dan R&D*.
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch Pada Assessment Pendidikan* (I). Trim Komunikata. [https://www.researchgate.net/publication/282673464\\_Aplikasi\\_Pemodelan\\_Rasch\\_Pada\\_Assessment\\_Pendidikan](https://www.researchgate.net/publication/282673464_Aplikasi_Pemodelan_Rasch_Pada_Assessment_Pendidikan)
- Tenny, Khairun Nisa, A., & Murtaolah. (2021). *Pengembangan Literasi Dan Numerasi Dalam Proses Belajar Dan Mengajar Berbagai Mata Pelajaran* (F. Mandik & T. Nira Sari, Eds.). Holifatul, S. (2021). Analisis Kemampuan Literasi Numerasi Ditinjau Dari Pengetahuan Metakognitif Siswa Dalam Menyelesaikan Soal Pisa Konten Space And Shape. Surabaya: Uin Sunan Ampel. <https://repository.kemendikdasmen.go.id/29935/1/Pengembangan%20literasi%20dan%20numerasi%20dalam%20proses%20belajar%20dan%20mengajar%20berbagai%20mata%20pelajaran.Pdf>
- Tessmer, M. (1993). *Planning And Conducting Formative Evaluations*. Routledge.
- Utari, T., Susanta, A., & Koto, I. (2024). *Pengembangan Instrumen Asesmen Kompetensi Minimum (Akm) Numerasi Model Pisa Dengan Konteks Musi Rawas Kelas V Sekolah Dasar*. 3(1). <https://ejournal.unib.ac.id/index.php/kapedas/index>
- Yandi, A., Nathania, A., Putri, K., & Syaza, Y. (2023). *Faktor-Faktor Yang Mempengaruhi Hasil Belajar Peserta Didik (Literature Review)*. <https://doi.org/10.38035/jpsn.v1i1>
- Zulfia, R. (2022). *Pengembangan Soal Akm (Asesmen Kompetensi Minimum) Numerasi Untuk Siswa Smp*. Universitas Islam Negeri Ar-Raniry Banda Aceh.