

PENGEMBANGAN INSTRUMEN DIAGNOSTIK LITERASI MATEMATIKA UNTUK ANALISIS PROFIL KOGNITIF SISWA SMP PADA MATERI SPLDV

Nugro Krismanto^{1*}, Wardani Rahayu², Achmad Ridwan³

^{1,2,3}Program Studi Penelitian dan Evaluasi Pendidikan, Universitas Negeri Jakarta, Jakarta, Indonesia

*Penulis korespondensi: nugro.krismanto@bpkpenaburjakarta.or.id

ABSTRAK

Hasil *Programme for International Student Assessment* (PISA) 2022 menunjukkan bahwa kemampuan matematika siswa Indonesia masih berada di bawah rata-rata internasional, khususnya pada materi aljabar yang menuntut kemampuan pemodelan dan penalaran. Salah satu materi krusial pada jenjang SMP adalah Sistem Persamaan Linear Dua Variabel (SPLDV), yang sering menjadi sumber kesulitan belajar siswa. Penelitian ini bertujuan untuk mengembangkan instrumen diagnostik guna menganalisis profil penguasaan atribut kognitif siswa SMP pada materi SPLDV menggunakan pendekatan *Cognitive Diagnostic Model* (CDM). Instrumen dikembangkan dalam bentuk 15 butir soal *Pilihan Ganda Kompleks* (PGK) yang mengukur empat atribut kognitif dan divalidasi oleh lima pakar. Validitas isi dianalisis menggunakan koefisien *Aiken's V*, sedangkan kualitas butir dan reliabilitas dianalisis melalui *Classical Test Theory* (CTT) dan *Item Response Theory* (IRT) menggunakan model *Generalized Partial Credit Model* (GPCM), termasuk evaluasi kecocokan model, karakteristik respons butir, serta *Test Information Function* (TIF) dan *Standard Error of Measurement* (SEM). Data selanjutnya dianalisis menggunakan model *GDINA* untuk menghasilkan profil penguasaan atribut siswa. Uji coba awal melibatkan 200 siswa, diikuti uji coba utama pada 756 siswa kelas VIII dari tiga provinsi. Hasil analisis menunjukkan bahwa sebagian besar butir memenuhi kriteria kualitas psikometrik dan asumsi pengukuran, dengan validitas isi yang baik (*Aiken's V* $\geq 0,80$) serta reliabilitas tinggi ($\alpha = 0,829$). Berdasarkan analisis CTT, IRT, dan CDM, dua butir dieliminasi sehingga diperoleh instrumen final berjumlah 13 butir. Instrumen ini dinyatakan valid, reliabel, dan efektif untuk mendiagnosis kesulitan belajar serta memetakan profil penguasaan atribut kognitif siswa pada materi SPLDV.

Kata kunci: *Cognitive Diagnostic Model* (CDM), *GDINA*, *Item Response Theory* (IRT), *Generalized Partial Credit Model* (GPCM), Sistem Persamaan Linear Dua Variabel (SPLDV)

1 PENDAHULUAN

Tantangan utama pendidikan matematika Indonesia saat ini adalah mengatasi kesenjangan kompetensi global yang tercermin dari hasil asesmen PISA. Hasil PISA 2022 menunjukkan bahwa kemampuan matematika siswa Indonesia berada jauh di bawah rata-rata global, dengan skor rata-rata hanya 366 poin (OECD, 2023). Data tersebut mengindikasikan bahwa hanya sebagian kecil siswa (sekitar 18%) yang mencapai tingkat kompetensi Level 2, yaitu batas kemampuan fungsional minimum. Kondisi ini menegaskan perlunya instrumen evaluasi yang mampu mendiagnosis secara spesifik letak kesulitan belajar, bukan hanya mengukur capaian akhir.

Materi Sistem Persamaan Linear Dua Variabel (SPLDV) merupakan fondasi krusial dalam aljabar tingkat menengah. Kesulitan siswa dalam materi ini seringkali melibatkan kemampuan pemodelan matematis dan interpretasi hasil, yang merupakan inti dari literasi matematika (OECD, 2019; Kholid & Aditia, 2023). Penilaian konvensional (CTT) gagal memberikan informasi rinci mengenai pola penguasaan atribut kognitif spesifik yang mendasari kinerja siswa pada SPLDV (Rizki & Retnawati, 2020).

Cognitive Diagnostic Model (CDM) menawarkan solusi evaluasi yang superior. CDM berfokus pada identifikasi atribut kognitif diskret (pengetahuan, keterampilan, strategi) yang dibutuhkan siswa untuk menyelesaikan butir soal, menghasilkan profil penguasaan biner (0/1) untuk setiap atribut (Maradani & Wilson, 2021). Oleh karena itu, penelitian ini bertujuan utama untuk: 1) Mengembangkan instrumen diagnostik literasi matematika pada materi SPLDV dengan format Pilihan Ganda Kompleks (PGK) yang mengukur empat atribut kognitif, dan 2) Menganalisis validitas dan reliabilitas instrumen tersebut menggunakan kombinasi pendekatan Aiken's V, CTT, IRT dan CDM-GDINA.

2 METODE

2.1 Desain dan Subjek Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan penekanan pada pengembangan dan validasi instrumen diagnostik berbasis teori pengukuran modern. Proses pengembangan didasarkan pada prinsip bahwa instrumen harus memiliki kejelasan konstruk, indikator terukur, dan teknik validasi yang kuat (DeVellis, 2017; Cohen et al., 2018). Instrumen dirancang untuk mengidentifikasi penguasaan atribut kognitif siswa pada materi Sistem Persamaan Linear Dua Variabel (SPLDV), yang merupakan salah satu kompetensi literasi matematika. Struktur instrumen dievaluasi melalui validitas isi, analisis empiris berdasarkan Classical Test Theory (CTT), *Item Response Theory* (IRT), serta *Cognitive Diagnostic Models* (CDM) untuk menghasilkan penilaian diagnostik yang lebih presisi (Maradani & Wilson, 2021; Rizki & Retnawati, 2020). Seluruh analisis statistik dan pemodelan dilakukan menggunakan program R dengan paket yang relevan untuk analisis CTT, IRT, *Generalized Partial Credit Model* (GPCM), dan CDM-GDINA.

Subjek penelitian terdiri atas lima pakar pendidikan matematika dan asesmen yang terlibat dalam proses validitas isi untuk memastikan kesesuaian konstruk dan kualitas butir instrumen (Retnawati, 2016). Uji coba awal melibatkan 200 siswa kelas VIII dari dua SMP di Jakarta dan Bekasi untuk menganalisis reliabilitas dan karakteristik butir menggunakan pendekatan Classical Test Theory (Cohen et al., 2018). Tahap ini bertujuan mengidentifikasi butir bermasalah sebelum dilakukan pengujian lanjutan. Selanjutnya, uji coba utama melibatkan 756 siswa kelas VIII dari tiga provinsi—Jakarta, Jawa Barat, dan Banten—untuk analisis Item Response Theory dan Cognitive Diagnostic Models. Sampel dipilih melalui teknik purposive cluster sampling guna memastikan representasi kemampuan siswa yang beragam (Anwar & Kartika, 2023; Auliya, 2022).

Pemilihan subjek tersebut memastikan bahwa proses pengembangan instrumen mencakup evaluasi teoretis dan empiris secara menyeluruh, sebagaimana direkomendasikan dalam pengembangan instrumen pendidikan modern (DeVellis, 2017; Cohen et al., 2018). Kombinasi data kualitatif dan kuantitatif memberikan dasar yang kuat untuk menghasilkan instrumen diagnostik yang valid dan reliabel serta mampu memberikan informasi diagnostik yang mendalam mengenai kemampuan siswa pada materi SPLDV, sejalan dengan prinsip analisis diagnostik berbasis model kognitif (Rizki & Retnawati, 2020; Maradani & Wilson, 2021).

2.2 Tahapan Pengembangan Instrumen

2.2.1 Identifikasi Konstruk dan Penyusunan Atribut

Identifikasi konstruk literasi matematika pada materi Sistem Persamaan Linear Dua Variabel (SPLDV) dalam penelitian ini mengacu pada kerangka asesmen PISA yang menekankan kemampuan memahami masalah kontekstual, memodelkan situasi matematis, menerapkan

prosedur penyelesaian, serta mengevaluasi dan menafsirkan solusi dalam konteks nyata (OECD, 2019). Kerangka tersebut selaras dengan Capaian Pembelajaran (CP) Fase D Kurikulum Merdeka, yang menuntut siswa mampu memformulasikan masalah kontekstual, menyelesaikan SPLDV secara tepat, dan menginterpretasikan hasilnya (Kemendikbudristek, 2022).

Berdasarkan kerangka tersebut, penelitian ini menggunakan empat atribut kognitif SPLDV yang telah divalidasi oleh pakar, yaitu: (A1) mengidentifikasi SPLDV; (A2) memodelkan masalah kontekstual ke dalam bentuk SPLDV; (A3) menyelesaikan SPLDV dengan metode yang sesuai; dan (A4) mengevaluasi serta menafsirkan solusi dalam konteks permasalahan nyata. Perumusan atribut mengikuti prinsip construct clarity untuk memastikan kejelasan konstruk dan konsistensi pengukuran (DeVellis, 2017). Keempat atribut dipetakan ke dalam Q-matrix sebagai dasar analisis Cognitive Diagnostic Model (CDM), yang ketepatannya menentukan akurasi inferensi profil penguasaan atribut siswa (Maradani & Wilson, 2021; Rizki & Retnawati, 2020).

Secara konseptual, atribut SPLDV memiliki hubungan hierarkis, namun analisis diagnostik menggunakan model GDINA yang memperlakukan atribut secara non-hierarkis untuk tujuan diagnostik eksploratif. Pendekatan ini memungkinkan estimasi penguasaan atribut secara empiris, sementara hubungan hierarkis tetap digunakan sebagai dasar interpretasi hasil (de la Torre, 2011; OECD, 2019).

2.2.2 Penyusunan Butir Soal Pilihan Ganda Kompleks (PGK)

Penyusunan butir instrumen dilakukan dengan mengembangkan 15 soal Pilihan Ganda Kompleks (PGK) yang disesuaikan dengan konteks literasi matematika dan karakteristik kompetensi SPLDV. Format PGK dipilih karena mampu mengungkap proses berpikir siswa secara lebih komprehensif, khususnya pada kemampuan berpikir tingkat tinggi (Higher Order Thinking Skills), serta memungkinkan pendeteksian respons parsial yang tidak dapat diperoleh melalui pilihan ganda konvensional (Amelia, 2016; Anjarrani & Kurniasih, 2023).

Setiap butir dikembangkan berdasarkan indikator atribut yang telah ditetapkan dengan memperhatikan kesesuaian konteks, representasi matematis, dan variasi tingkat kognitif. Konteks autentik digunakan untuk menilai kemampuan siswa dalam memahami, memodelkan, dan menyelesaikan permasalahan nyata, sementara setiap opsi jawaban dirancang untuk mengidentifikasi pola kesalahan konseptual siswa.

2.2.3 Validitas Isi

Validitas isi instrumen dievaluasi oleh lima pakar di bidang pendidikan matematika dan evaluasi pembelajaran menggunakan koefisien Aiken's V untuk menilai tingkat kesepakatan terhadap relevansi butir instrumen (Yusrizal et al., 2020). Penilaian mencakup kesesuaian indikator dengan konstruk, relevansi konteks soal, kejelasan representasi matematis, serta ketepatan pemetaan atribut dalam Q-matrix. Butir dinyatakan memiliki validitas isi memadai apabila nilai Aiken's $V \geq 0,80$ (Retnawati, 2016). Butir yang tidak memenuhi kriteria, direvisi atau dieliminasi agar hanya butir dengan kesesuaian teoretis tinggi yang digunakan pada tahap uji coba empiris, sehingga menjamin keterpaduan antara konstruk teoretis dan bukti empiris dalam pengembangan instrumen diagnostik.

2.3 Uji Coba dan Analisis Psikometrik

2.3.1 Analisis *Classical Test Theory* (CTT)

Uji coba awal dengan 200 siswa dianalisis menggunakan CTT untuk memperoleh estimasi reliabilitas, tingkat kesukaran, dan daya pembeda (Cohen et al., 2018). Reliabilitas dihitung

menggunakan koefisien Alpha Cronbach, yang merupakan metode umum dalam penelitian pendidikan (Retnawati, 2016). Butir yang tidak memenuhi kriteria statistik direvisi atau dieliminasi.

2.3.2 Uji Kecocokan Model dan Independensi Lokal.

Kecocokan model IRT dievaluasi menggunakan indeks M2 dan RMSEA untuk memastikan kesesuaian pola respons siswa dengan model yang diuji. Penelitian ini mempertimbangkan tiga model politomus, yaitu Graded Response Model (GRM) dan Generalized Partial Credit Model (GPCM). GRM digunakan untuk respons ordinal kumulatif, sedangkan GPCM sesuai untuk skor parsial; GPCM dipilih karena fleksibilitasnya dalam mengestimasi parameter ambang dan diskriminasi pada butir PGK (Cohen et al., 2018; Maradani & Wilson, 2021). Asumsi independensi lokal diperiksa menggunakan residual Q3 untuk mendeteksi korelasi antarbutir yang tidak diharapkan, di mana nilai Q3 tinggi mengindikasikan ketergantungan lokal yang dapat mengganggu validitas estimasi parameter (DeVellis, 2017). Kombinasi indeks M2, RMSEA, dan Q3 memberikan dasar yang memadai dalam menentukan model IRT yang paling sesuai dengan data respons siswa.

2.3.3 Analisis IRT Menggunakan Model GPCM

Uji coba lanjutan dianalisis menggunakan *Generalized Partial Credit Model* (GPCM), karena model ini sesuai untuk butir politomus seperti PGK (Maradani & Wilson, 2021). Analisis mencakup parameter diskriminasi dan threshold, serta pemeriksaan item fit, fungsi informasi butir (IIF), fungsi informasi tes (TIF), dan kesalahan pengukuran (SEM). Keakuratan pengukuran dievaluasi pada tingkat kepercayaan 95%.

2.3.4 Analisa CDM Menggunakan GDINA

Model GDINA digunakan untuk menghasilkan profil atribut siswa melalui estimasi pola penguasaan keterampilan. Pendekatan ini memungkinkan pemetaan kemampuan siswa secara lebih rinci dan sesuai dengan tujuan pengukuran diagnostik (Rizki & Retnawati, 2020). Hasil CDM memberikan proporsi penguasaan atribut, attribute mastery profile, dan evaluasi Q-matrix.

3 HASIL DAN PEMBAHASAN

3.1 Penyusunan Atribut

Validitas konstruk instrumen diagnostik ini dijamin melalui perancangan atribut kognitif yang selaras dengan dimensi Literasi Matematika PISA dan Capaian Pembelajaran (CP) Fase D Kurikulum Merdeka. Empat atribut (A1–A4) merepresentasikan tahapan hierarkis pemecahan masalah matematika (OECD, 2019). Atribut A1 (mengidentifikasi SPLDV) dan A2 (menyusun model SPLDV) mencerminkan tahap formulate, yang menuntut kemampuan mentransformasikan konteks masalah ke dalam model aljabar dan sejalan dengan tuntutan CP Fase D Elemen Aljabar (Kholid & Aditia, 2023). Atribut A3 (menyelesaikan SPLDV) merepresentasikan tahap employ dan reasoning, yang menekankan kemampuan menerapkan serta menjustifikasi prosedur penyelesaian secara efisien (Schoenfeld, 1985). Atribut A4 (mengevaluasi solusi SPLDV) merupakan tahap kognitif tertinggi yang mencakup interpretasi, refleksi, dan penalaran tingkat lanjut, sehingga relevan untuk mengukur kompetensi di luar ranah prosedural dan sesuai dengan tuntutan Kurikulum Merdeka.

Koherensi konseptual tersebut dioperasionalkan melalui Q-matrix yang menetapkan hubungan eksplisit antara butir dan atribut (Rizki & Retnawati, 2020). Q-matrix yang valid merupakan prasyarat utama dalam pemodelan Cognitive Diagnostic Model berbasis GDINA, karena memungkinkan estimasi profil penguasaan atribut siswa secara akurat dan tidak bias, serta mendukung diagnosis kesulitan belajar yang informatif untuk intervensi pembelajaran yang terarah (de la Torre, 2011).

3.2 Validitas Isi Instrumen

Hasil validitas isi menunjukkan bahwa keempat atribut kognitif—A1 (memahami konsep), A2 (membuat model matematika), A3 (menyelesaikan SPLDV), dan A4 (menginterpretasi solusi)—memiliki nilai Aiken's V pada rentang 0,80–1,00. Temuan ini menegaskan bahwa struktur atribut layak secara teoretis dan representatif untuk mengukur kompetensi SPLDV (DeVellis, 2017; OECD, 2019).

Validitas isi instrumen diperoleh melalui penilaian lima pakar menggunakan koefisien Aiken's V terhadap 15 butir soal. Seluruh butir (B1–B15) dinyatakan valid dengan nilai V antara 0,73–1,00, melampaui batas minimum $V \geq 0,70$ yang direkomendasikan dalam pengembangan instrumen pendidikan (Retnawati, 2016). Beberapa butir (B2, B7, B8, dan B10) mencapai nilai $V = 1,00$, menunjukkan kesepakatan pakar yang sempurna, sementara butir multiatribut seperti B12 ($V = 0,93$) dan B15 ($V = 0,87$) juga menunjukkan kualitas konten yang kuat. Butir dengan nilai terendah (B1 dan B9; $V = 0,73$) tetap memenuhi kriteria dan layak diuji secara empiris.

Penilaian pakar terhadap keterkaitan atribut–butir menunjukkan kesesuaian yang konsisten antara tuntutan kognitif dan konstruk yang diukur. Seluruh pemetaan atribut dalam Q-matrix memperoleh nilai Aiken's $V \geq 0,80$, menegaskan bahwa Q-matrix valid dan layak digunakan untuk analisis diagnostik berbasis CDM (Retnawati, 2016; Yusrizal et al., 2020).

3.3 Hasil Analisis Classical Test Theory (CTT)

3.3.1 Reliabilitas Tes

Tabel 1. Reliabilitas Instrumen

N	Mean Total Score	SD Total Score	Ave_r	SD_r	Alpha
200	33,731	10,432	0,236	0,169	0,829

Keterangan: N = jumlah responden; Mean Total Score = rerata skor total tes; SD Total Score = simpangan baku skor total tes; Ave r = rerata korelasi antarbutir; SD r = simpangan baku korelasi antarbutir; Alpha = koefisien reliabilitas Alpha Cronbach.

Sumber: (Krismanto *et al.*, 2025)

Reliabilitas instrumen berdasarkan uji coba pada 200 siswa SMP menunjukkan nilai Alpha Cronbach sebesar 0,829, yang mengindikasikan konsistensi internal yang kuat (Nunnally & Bernstein, 1994). Nilai ini menunjukkan bahwa instrumen layak digunakan pada analisis lanjutan, termasuk pemodelan IRT dan diagnostik GDINA.

Namun, mengingat instrumen ini bersifat multidimensi dan mengukur empat atribut kognitif, Alpha Cronbach digunakan sebagai indikator awal konsistensi internal, bukan sebagai bukti definitif reliabilitas. Dalam konteks ini, nilai Alpha merefleksikan korelasi rata-rata antarbutir dan berpotensi mengestimasi reliabilitas secara berlebihan, sehingga interpretasinya dilengkapi dengan analisis psikometrik lanjutan.

3.3.2 Indeks Kesukaran dan daya Beda

Analisis kualitas butir dilakukan menggunakan indeks kesukaran (p) dan daya beda (pBis) berdasarkan Classical Test Theory (CTT). Indeks kesukaran merefleksikan proporsi siswa yang menjawab benar, sedangkan daya beda menunjukkan kemampuan butir dalam membedakan siswa berkemampuan tinggi dan rendah (Allen & Yen, 2002). Sebagian besar butir berada pada rentang kesukaran optimal (0,30–0,80), sehingga dinilai fungsional untuk mengukur variasi kemampuan siswa pada materi SPLDV (Nunnally & Bernstein, 1994).

Daya beda dievaluasi melalui korelasi butir–total terkoreksi (pBis). Butir dengan pBis $\geq 0,30$ (misalnya B1, B3, B8, B11, dan B15) menunjukkan kemampuan diskriminatif yang memadai. Sebaliknya, butir B7 dan B13 dieliminasi karena memiliki daya beda rendah atau negatif. Eliminasi ini penting untuk menjaga asumsi monotonisitas dan mencegah bias estimasi parameter pada analisis IRT dan CDM–GDINA, sehingga data final memiliki kualitas respons yang lebih stabil dan andal.

3.4 Hasil Analisis Item Respon Theory (IRT)

3.4.1 Uji Kecocokan Model

Tabel 2. Ringkasan Uji Kecocokan Model

Model	AIC	SABIC	HQ	BIC	logLik	X ²	df	p
PCM	27923.60	28009.31	28028.77	28196.65	- 13902.80	-	-	-
GPCM	27501.43	27607.48	27631.56	27839.28	- 13677.72	450.166	14	0
GRM	27498.89	27604.93	27629.02	27836.73	- 13676.44	2.547	0	NA

Keterangan: PCM = Partial Credit Model; GPCM = Generalized Partial Credit Model; GRM = Graded Response Model; AIC = Akaike Information Criterion; SABIC = Sample-Size Adjusted Bayesian Information Criterion; HQ = Hannan–Quinn Information Criterion; BIC = Bayesian Information Criterion; logLik = log-likelihood; χ^2 = statistik uji kecocokan; df = derajat kebebasan; p = nilai signifikansi; NA = not available; (–) = tidak dihitung. Sumber: (Krismanto *et al.*, 2025)

Uji kecocokan model dilakukan dengan membandingkan Partial Credit Model (PCM), Generalized Partial Credit Model (GPCM), dan Graded Response Model (GRM). Meskipun GRM menunjukkan nilai AIC terendah, model ini mensyaratkan struktur respons ordinal kumulatif yang tidak sesuai dengan karakter penskoran parsial pada butir Pilihan Ganda Kompleks (PGK). Sebaliknya, GPCM menunjukkan kecocokan model yang baik dan lebih tepat secara konseptual karena mampu menangani respons politomus non-ordinal serta memberikan estimasi parameter yang lebih fleksibel dibandingkan PCM. Oleh karena itu, GPCM dipilih sebagai model final dalam analisis IRT (Cohen *et al.*, 2018; Maradani & Wilson, 2021).

Pemilihan GPCM didasarkan pada karakter penskoran PGK, di mana kategori skor (0–4) merepresentasikan kombinasi penguasaan atribut kognitif, bukan tingkatan kemampuan yang bersifat ordinal. Dengan demikian, transisi antar kategori skor tidak mencerminkan peningkatan kemampuan yang seragam, melainkan penguasaan set atribut yang berbeda. GPCM tidak mensyaratkan asumsi ordinalitas kategori respons dan hanya mengaitkan kategori skor dengan tingkat kesulitan item, sehingga lebih sesuai untuk instrumen dengan penskoran parsial berbasis tahapan penyelesaian dan penguasaan atribut kognitif yang berbeda.

3.4.2 Uji Asumsi dan Independensi Lokal

Tabel 3. Ringkasan Uji Asumsi

Item A	Item B	r	partial-r	Q3	χ^2	p- χ^2	G ²	p-G ²	JSI	Keterangan
B11	B14	0.798	0.568	0.565	995.382	0.101	4.853	0.907	1.071	Tidak terpenuhi

Keterangan: Item A–Item B = pasangan butir yang diuji; r = korelasi antarbutir; *partial-r* = korelasi parsial setelah mengendalikan kemampuan laten (θ); $Q3$ = residual korelasi Yen (uji *local independence*); χ^2 = statistik *Chi-Square*; $p-\chi^2$ = nilai signifikansi *Chi-Square*; G^2 = statistik *likelihood ratio*; $p-G^2$ = nilai signifikansi G^2 ; JSI = *Jackknife Stability Index*. Sumber: (Krismanto *et al.*, 2025)

Dari 105 pasangan butir yang diuji, hanya satu pasangan (0,95%) yang melampaui batas residual $Q3 > 0,30$, jauh di bawah ambang toleransi umum sebesar 5%. Temuan ini menunjukkan bahwa asumsi independensi lokal terpenuhi, sehingga estimasi parameter dalam model IRT dapat dianggap stabil dan analisis lanjutan dapat dilakukan tanpa koreksi residual tambahan (Cohen *et al.*, 2018). Residual $Q3$ digunakan sebagai indikator pelanggaran ketergantungan lokal, di mana nilai $Q3$ yang tinggi merefleksikan korelasi residual antarbutir yang tidak diharapkan (Maradani & Wilson, 2021). Rendahnya tingkat pelanggaran pada instrumen ini mengindikasikan bahwa korelasi antarbutir sebagian besar dijelaskan oleh kemampuan laten (θ) atau profil atribut yang diukur, sehingga struktur butir dinilai konsisten dan estimasi parameter item maupun profil atribut dalam analisis IRT dan CDM–GDINA bersifat valid dan tidak bias.

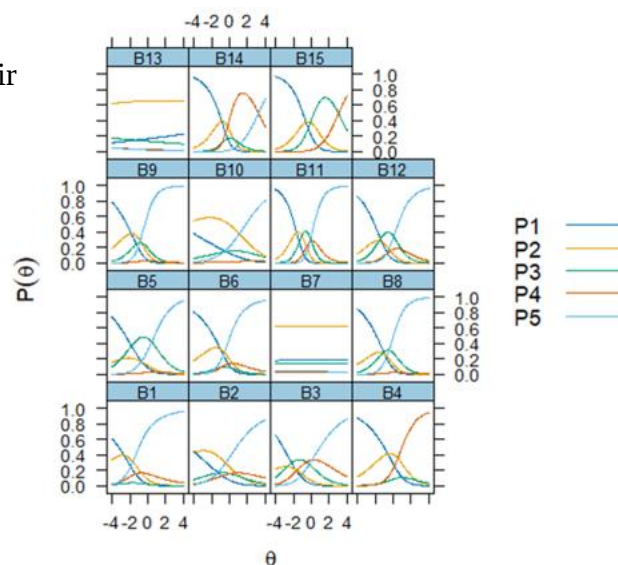
3.4.3 Uji Parameter Butir (Daya Beda dan Tingkat Kesukaran)

Analisis indeks kesukaran menunjukkan bahwa sebagian besar butir berada dalam rentang ideal 0,30–0,80, sehingga tidak terlalu mudah maupun terlalu sulit untuk mengukur variasi kemampuan siswa pada materi SPLDV (Nunnally & Bernstein, 1994). Daya beda butir juga menunjukkan kualitas yang baik, dengan beberapa butir (B1, B3, B8, B11, dan B15) memiliki nilai $pBis \geq 0,30$, yang menandakan kemampuan diskriminatif yang memadai dalam membedakan siswa berkemampuan tinggi dan rendah (Allen & Yen, 2002).

Butir B7 dan B13 memiliki daya beda rendah hingga negatif, mengindikasikan potensi ambiguitas atau ketidaktepatan kunci jawaban, sehingga dieliminasi untuk menjaga asumsi monotonisitas dan mencegah bias estimasi parameter pada analisis IRT dan CDM–GDINA (de la Torre, 2011). Eliminasi ini memastikan bahwa data final yang dianalisis memiliki kualitas respons yang lebih stabil dan andal.

Secara keseluruhan, pemenuhan kriteria kualitas CTT berfungsi sebagai tahap screening awal item. Instrumen hasil seleksi yang terdiri dari 13 butir final menjadi prasyarat metodologis untuk menjamin stabilitas data respons dan akurasi estimasi parameter laten pada analisis IRT dan diagnostik GDINA.

3.4.4 Fungsi Probabilitas Butir

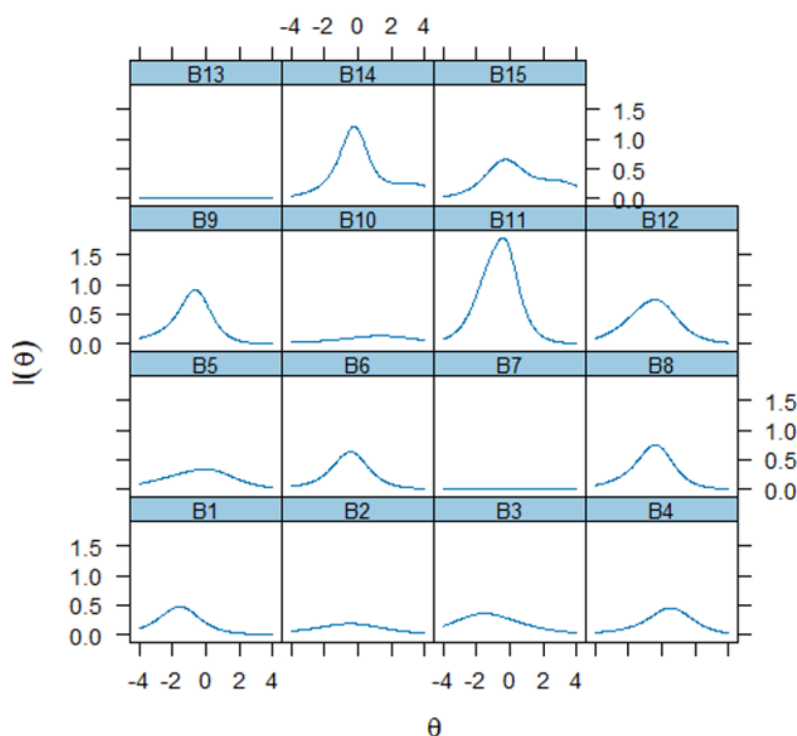


Gambar 1. Fungsi Probabilitas Butir (Krismanto *et al.*, 2025)

Analisis *Item Probability Functions* (IPF) menunjukkan bahwa sebagian besar butir memenuhi asumsi monotonisitas, yaitu probabilitas memperoleh kategori skor yang lebih tinggi meningkat seiring peningkatan kemampuan laten. Dalam konteks Generalized Partial Credit Model (GPCM), pemenuhan asumsi ini penting untuk memastikan bahwa transisi antar kategori skor merefleksikan penguasaan atribut yang semakin kompleks.

Butir B7 dan B13 menunjukkan pola respons non-monotonik dengan kurva probabilitas yang tidak teratur. Pelanggaran ini mengindikasikan ketidakmampuan butir dalam merepresentasikan hierarki atribut yang diharapkan ($A1 > A2 > A3 > A4$), yang berpotensi disebabkan oleh ketidaksesuaian antara kunci jawaban dan pemetaan atribut dalam Q-matrix atau adanya cue yang memungkinkan respons benar tanpa penguasaan atribut tingkat tinggi. Oleh karena itu, kedua butir tersebut dinilai tidak layak secara diagnostik.

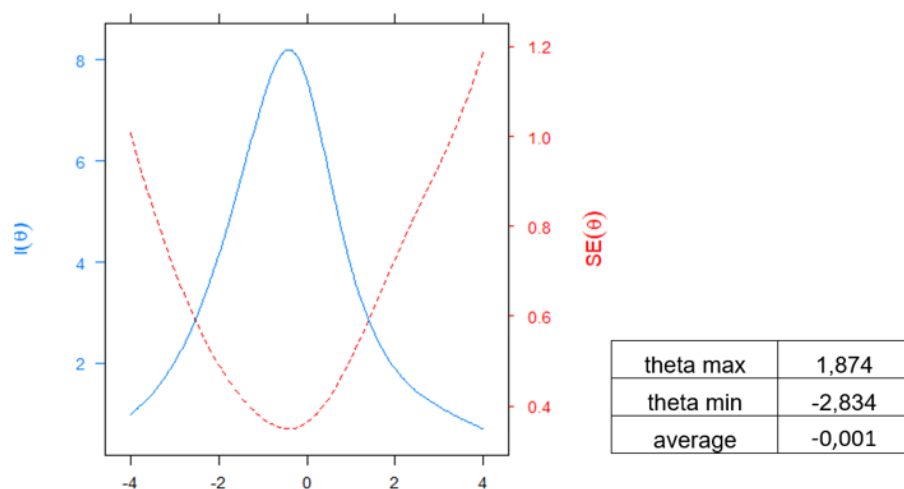
3.4.5 Analisis Fungsi Informasi Butir



Gambar 2. Fungsi Informasi Butir (Krismananto *et al.*,2025)

Analisis *Item Probability Functions* (IPF) menunjukkan bahwa sebagian besar butir memenuhi asumsi monotonisitas, di mana probabilitas memperoleh kategori skor yang lebih tinggi meningkat seiring peningkatan kemampuan laten, sebagaimana disyaratkan dalam Generalized Partial Credit Model (GPCM). Sebaliknya, butir B7 dan B13 memperlihatkan pola respons non-monotonik yang tidak merepresentasikan hierarki atribut yang diharapkan ($A1 > A2 > A3 > A4$), kemungkinan akibat ketidaksesuaian pemetaan atribut atau adanya cue jawaban sehingga dinilai tidak layak secara diagnostik.

3.4.6 Analisis Informasi Tes



Gambar 3. Fungsi Informasi Tes (Krismanto *et al.*, 2025)

Analisis *Test Information Function* (TIF) menunjukkan bahwa instrumen bekerja optimal untuk tujuan diagnostik, dengan puncak informasi pada sekitar $\theta = -0,40$, yang menandakan presisi pengukuran tertinggi berada pada tingkat kemampuan sedikit di bawah rata-rata. Pada kemampuan sangat rendah atau sangat tinggi, nilai informasi menurun dan *Standard Error of Measurement* (SEM) meningkat, sebagaimana karakteristik instrumen dengan rentang kesulitan moderat (Cohen *et al.*, 2018). Pola ini selaras dengan tujuan penelitian, yaitu mengidentifikasi kesulitan belajar, karena presisi tertinggi berada pada rentang kemampuan rendah–sedang.

Peningkatan SEM pada $\theta < -2,0$ dan $\theta > 1,5$ merefleksikan batas presisi pengukuran, namun tidak mengurangi validitas studi karena fokus utama instrumen adalah diagnosis defisit atribut pada mayoritas siswa. Distribusi kemampuan siswa ($\theta = -2,834$ hingga $1,874$; rata-rata $\theta \approx 0$) menunjukkan variasi kemampuan yang memadai, dan kedekatan puncak TIF dengan pusat distribusi menegaskan bahwa instrumen dirancang efektif untuk memberikan informasi diagnostik yang relevan bagi siswa SMP pada materi SPLDV.

3.5 Q-Matrik yang Digunakan untuk Pengukuran G DINA

Penyusunan Q-matrix dilakukan untuk memastikan keselarasan setiap butir dengan atribut kognitif SPLDV yang diukur. Q-matrix merupakan komponen kunci dalam analisis diagnostik karena menetapkan hubungan eksplisit antara butir dan keterampilan yang diperlukan untuk menyelesaikannya, sekaligus menjamin validitas konstruk dalam model diagnostik kognitif (Rizki & Retnawati, 2020).

Empat atribut dalam instrumen ini selaras dengan tahapan Literasi Matematika PISA (OECD, 2019): A1 dan A2 merepresentasikan proses *formulate* (identifikasi variabel dan pemodelan SPLDV dari konteks masalah), A3 merepresentasikan proses *employ* (penerapan prosedur penyelesaian SPLDV), dan A4 merepresentasikan proses *interpret and reflect* (penilaian dan penafsiran solusi dalam konteks nyata). Integrasi ini memastikan bahwa instrumen mengukur tidak hanya penguasaan konten SPLDV, tetapi juga dimensi kognitif yang kompleks dan bermakna.

3.5 Interpretasi Secara Keseluruhan

Berdasarkan hasil analisis validitas dan kualitas pengukuran, instrumen diagnostik SPLDV yang dikembangkan menghasilkan bentuk final yang terdiri dari 13 butir soal. Seluruh butir memenuhi kriteria validitas isi dengan nilai *Aiken's V* $\geq 0,80$, yang menunjukkan bahwa para pakar sepakat bahwa setiap butir soal telah sesuai dengan tujuan pengukuran dan benar-benar merepresentasikan atribut kognitif SPLDV yang ingin dianalisis. Dengan kata lain, instrumen ini secara konseptual sudah tepat sasaran, karena butir-butirnya mengukur kemampuan yang memang ingin diukur, bukan kemampuan lain di luar SPLDV.

Hasil analisis reliabilitas menunjukkan koefisien $\alpha = 0,829$, yang mengindikasikan konsistensi internal yang tinggi. Artinya, apabila instrumen ini digunakan pada kelompok siswa yang sebanding, hasil pengukuran yang diperoleh cenderung stabil dan tidak bersifat kebetulan. Konsistensi ini penting dalam konteks diagnosis, karena keputusan pembelajaran yang diambil guru perlu didasarkan pada hasil pengukuran yang dapat dipercaya.

Analisis menggunakan *Classical Test Theory* (CTT) menunjukkan bahwa sebagian besar butir memiliki tingkat kesukaran sedang dan daya beda $\geq 0,30$. Tingkat kesukaran sedang menandakan bahwa soal tidak terlalu mudah maupun terlalu sulit, sehingga mampu mengungkap variasi kemampuan siswa. Daya beda yang memadai menunjukkan bahwa butir-butir soal dapat membedakan siswa dengan tingkat penguasaan SPLDV yang berbeda, sehingga informasi yang dihasilkan menjadi bermakna untuk tujuan diagnosis. Dua butir (B7 dan B13) dieliminasi karena tidak mampu membedakan kemampuan siswa secara jelas dan menunjukkan pola respons yang tidak konsisten. Keberadaan butir semacam ini berpotensi menyesatkan hasil diagnosis, sehingga penghapusan dilakukan untuk menjaga kualitas instrumen.

Selanjutnya, analisis *Item Response Theory* (IRT) menunjukkan bahwa *Generalized Partial Credit Model* (GPCM) merupakan model yang paling sesuai dengan karakteristik butir *Pilihan Ganda Kompleks*. Kesesuaian model ini menunjukkan bahwa skor parsial yang diberikan benar-benar mencerminkan tahapan penguasaan atribut kognitif siswa, bukan sekadar hasil benar atau salah. Informasi pengukuran paling tinggi berada pada rentang kemampuan rendah hingga sedang, yang berarti instrumen ini paling akurat ketika digunakan untuk mengidentifikasi siswa yang mengalami kesulitan belajar, sesuai dengan tujuan utama pengembangan instrumen diagnostik.

Secara keseluruhan, kombinasi hasil validitas isi, analisis *Classical Test Theory*, dan *Item Response Theory* menunjukkan bahwa instrumen ini tidak hanya layak secara statistik, tetapi juga mudah diinterpretasikan dan relevan untuk digunakan oleh guru sebagai alat bantu memahami letak kesulitan siswa dalam menyelesaikan SPLDV. Dengan demikian, hasil pengukuran dapat dimanfaatkan sebagai dasar perancangan intervensi pembelajaran yang lebih terarah dan sesuai dengan kebutuhan nyata siswa.

4 KESIMPULAN

Instrumen diagnostik Sistem Persamaan Linear Dua Variabel (SPLDV) yang dikembangkan dalam penelitian ini telah melalui proses pengembangan dan evaluasi yang komprehensif menggunakan kombinasi pendekatan *Aiken's V*, *Classical Test Theory* (CTT), *Item Response Theory* (IRT), dan *Cognitive Diagnostic Model* (CDM). Validitas isi yang dianalisis menggunakan *Aiken's V* menunjukkan bahwa seluruh atribut kognitif dan sebagian besar butir soal memiliki kesesuaian konstruk yang baik ($V \geq 0,80$). Analisis CTT menunjukkan bahwa instrumen memiliki reliabilitas tinggi ($\alpha = 0,829$), tingkat kesukaran moderat, serta daya beda yang memadai.

Selanjutnya, analisis IRT menggunakan *Generalized Partial Credit Model* (GPCM) menunjukkan kecocokan model yang baik dan informasi pengukuran yang stabil pada rentang

kemampuan rendah hingga sedang, yang sesuai dengan tujuan diagnostik instrumen. Berdasarkan evaluasi karakteristik respons butir dan asumsi pengukuran, dua butir (B7 dan B13) dieliminasi karena memiliki informasi sangat rendah dan pola respons yang tidak konsisten.

Analisis lanjutan menggunakan pendekatan CDM dengan model *GDINA* memungkinkan pemetaan profil penguasaan atribut kognitif siswa secara lebih rinci, sehingga instrumen tidak hanya berfungsi sebagai alat pengukuran hasil belajar, tetapi juga sebagai alat diagnosis kesulitan belajar siswa pada materi SPLDV. Dengan demikian, instrumen final yang terdiri dari 13 butir dinyatakan valid, reliabel, dan efektif untuk menganalisis penguasaan atribut kognitif siswa SMP.

Instrumen ini direkomendasikan sebagai alat diagnosis awal untuk memetakan kesulitan belajar siswa, khususnya pada aspek pemodelan dan interpretasi solusi SPLDV yang konsisten menjadi area kelemahan utama. Agar pemanfaatannya optimal, guru perlu dibekali kompetensi dalam menafsirkan hasil diagnostik berbasis CDM sehingga intervensi pembelajaran dapat dirancang secara lebih tepat sasaran. Penelitian selanjutnya disarankan mencakup penambahan butir pada tingkat kemampuan tinggi, pengujian pada sampel yang lebih luas, serta integrasi instrumen ke dalam platform digital untuk meningkatkan efisiensi pelaksanaan dan analisis asesmen.

UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih kepada dosen pembimbing atas bimbingan dan arahan akademik selama proses penelitian. Apresiasi juga diberikan kepada para validator ahli, pihak sekolah menengah pertama yang telah memberikan izin dan fasilitas penelitian, serta rekan mahasiswa dan rekan kerja atas dukungan ilmiah yang berkontribusi pada pelaksanaan studi. Ucapan terima kasih turut disampaikan kepada keluarga atas dukungan moral dan motivasi yang berkelanjutan hingga penelitian ini terselesaikan.

DAFTAR PUSTAKA

- Amelia, A. (2016). Pengembangan instrumen literasi numerasi berbasis HOTS pada materi aritmatika sosial. *Jurnal Riset Pendidikan Matematika*, 3(2), 112–125. <https://journal.unirow.ac.id>
- Anjarrani, A., & Kurniasih, R. (2023). Analisis kualitas soal numerasi dalam meningkatkan higher order thinking skills (HOTS) siswa. *Jurnal Pendidikan Matematika*, 10(2), 123–135.
- Anwar, F., & Kartika, R. (2023). Kemampuan literasi matematis calon guru dalam pengajuan masalah berorientasi data. *Jurnal Matematika dan Pendidikan Matematika*, 11(1), 65–78. <https://ojs.fkip.ummetro.ac.id/index.php/matematika/article/view/6463>
- Auliya, L. (2022). *Pengembangan instrumen tes literasi matematis berbasis budaya Betawi untuk meningkatkan pemahaman siswa dalam konteks budaya lokal* [Skripsi, Universitas Islam Negeri Syarif Hidayatullah Jakarta]. <https://repository.uinjkt.ac.id>
- Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). Routledge.
- DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Sage Publications.
- Kemendikbudristek. (2022). *Panduan pelaksanaan asesmen nasional*. Pusat Asesmen Pendidikan.
- Maradani, R., & Wilson, M. (2021). Development of diagnostic instruments in education: Applying the construct modeling approach. *Journal of Educational Measurement and Evaluation*, 15(2), 55–70.

- OECD. (2019). *PISA 2018 assessment and analytical framework: Mathematics, reading, science and global competence*. OECD Publishing.
- OECD. (2019). *PISA 2018 results: What students know and can do*. OECD Publishing. <https://doi.org/10.1787/5f07c754-en>
- OECD. (2021). *The future of education and skills: Assessment in the 21st century*. OECD Publishing.
- OECD. (2023). *Mathematics literacy and PISA performance: Trends and challenges in digital learning*. OECD Publishing.
- Retnawati, H. (2016). *Validitas dan reliabilitas instrumen penelitian*. Parama Publishing.
- Retnawati, H., Hadi, S., & Nugraha, A. (2016). Difficulties in the implementation of higher order thinking skills in mathematics learning curriculum. *Journal of Physics: Conference Series*, 012047.
- Rizki, S., & Retnawati, H. (2020). Developing diagnostic test instruments in mathematics education: A systematic approach. *Jurnal Riset Pendidikan Matematika*, 7(2), 145–160.
- Yusrizal, B., Susanti, T., & Astuti, R. (2020). Validasi instrumen dengan koefisien Aiken's V. *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia*, 9(3), 221–230.