

Evaluasi Psikometrik Instrumen Pengukuran Kemampuan Pemecahan Masalah Matematis Siswa berbasis Partial Credit Model

Psychometric Evaluation of a Mathematical Problem-Solving Instrument Using the Partial Credit Model

Hendarji Suryantoro*, Ilham Falani, Marlina

Program Studi Pendidikan Matematika, Universitas Jambi, Jambi, Indonesia

*Corresponding author: hendarjis@gmail.com

Abstract

Accurate measurement of students' mathematical problem-solving ability is a prerequisite for evaluating instructional quality, yet traditional assessments remain rooted in Classical Test Theory (CTT), which suffers from sample-dependency of parameters and lacks diagnostic information for partially correct responses. This study applied the Partial Credit Model (PCM), a polytomous Rasch family model, to evaluate the psychometric properties of an essay-type instrument designed to assess junior high school learners' competence in solving linear-equation systems (two-variable linear equations). The instrument was constructed according to Polya's four stages of problem solving, treating each stage as an ordered "item step." Using a quantitative descriptive approach, response data from students were analyzed with PCM to estimate item difficulty parameters, generate item characteristic curves, and examine content validity, reliability, item bias, standard error of measurement, and the test information function. Findings indicate that the instrument meets contemporary psychometric standards: the PCM provided precise threshold estimates between score categories, revealing specific cognitive bottlenecks within Polya's stages and offering richer diagnostic insight than total-score CTT analyses. Consequently, the developed test is both valid and reliable for competency-based assessment, and its PCM-based scoring is recommended for broader implementation in mathematics problem-solving evaluation.

Keywords: *Partial Credit Model, Rasch psychometrics, mathematical problem-solving, Polya's problem-solving steps, validity, reliability.*

Abstrak

Pengukuran yang akurat terhadap kemampuan pemecahan masalah matematis (KPPM) siswa merupakan prasyarat penting dalam menilai kualitas pembelajaran, namun penilaian tradisional masih banyak bergantung pada *Classical Test Theory* (CTT) yang memiliki kelemahan ketergantungan parameter pada sampel serta tidak memberikan informasi diagnostik untuk respons parsial. Penelitian ini menerapkan *Partial Credit Model* (PCM), sebuah model Rasch polytomous, untuk mengevaluasi sifat psikometrik instrumen esai yang dirancang mengukur kompetensi siswa SMP dalam menyelesaikan Sistem Persamaan Linear Dua Variabel (SPLDV). Instrumen disusun berdasarkan empat tahapan Polya, dengan masing-masing tahapan diperlakukan sebagai "item step" berurutan. Menggunakan pendekatan deskriptif kuantitatif, data respons siswa dianalisis dengan PCM guna memperkirakan parameter kesulitan butir, menghasilkan kurva karakteristik butir, serta menilai validitas isi, reliabilitas, bias butir, *Standard Error of Measurement* (SEM), dan fungsi informasi tes. Hasil menunjukkan bahwa instrumen memenuhi standar psikometri modern: PCM memberikan estimasi ambang batas yang tepat antar kategori skor, mengungkap titik kesulitan kognitif spesifik pada setiap tahapan Polya dan memberikan

wawasan diagnostik yang lebih kaya dibandingkan analisis total skor pada CTT. Dengan demikian, tes yang dikembangkan terbukti valid dan reliabel untuk penilaian berbasis kompetensi, serta skor berbasis PCM direkomendasikan untuk diterapkan secara lebih luas dalam evaluasi pemecahan masalah matematika.

Kata Kunci: Model Partial Credit, psikometri, Rasch model, tahapan Polya, reliabilitas

PENDAHULUAN

Kemampuan pemecahan masalah matematis (KPPM) merupakan kompetensi fundamental dalam pembelajaran matematika abad ke-21 karena berkaitan dengan kemampuan berpikir kritis, penalaran logis, kreativitas, dan pengambilan keputusan dalam situasi non-rutin. Sejalan dengan itu, National Council of Teachers of Mathematics menempatkan pemecahan masalah sebagai inti pembelajaran matematika untuk mengembangkan penalaran dan kemampuan berpikir tingkat tinggi siswa.

Urgensi KPPM semakin meningkat seiring implementasi pembelajaran berbasis kompetensi yang menuntut evaluasi kemampuan siswa secara lebih komprehensif. Pengukuran yang akurat diperlukan agar hasil evaluasi benar-benar merepresentasikan kemampuan siswa secara objektif (Faradila et al., 2025). Dalam perspektif psikometri modern, instrumen yang baik tidak hanya menghasilkan skor, tetapi juga memberikan informasi diagnostik mengenai karakteristik kemampuan peserta didik (Boone, 2016).

Namun, praktik evaluasi pembelajaran matematika di sekolah masih didominasi pendekatan berbasis Classical Test Theory (CTT). Pendekatan ini mudah diterapkan, tetapi memiliki keterbatasan mendasar karena kesukaran dan daya beda sangat dipengaruhi karakteristik sampel sehingga estimasinya berubah pada kelompok yang berbeda (Bond et al., 2021). Selain itu, CTT cenderung mereduksi kompleksitas proses berpikir siswa menjadi skor total

semata sehingga informasi diagnostik pada kemampuan siswa belum dapat dijelaskan secara rinci.

Keterbatasan tersebut semakin terlihat pada instrumen berbentuk soal esai pemecahan masalah matematis. Siswa dapat menunjukkan proses berpikir berbeda meskipun memperoleh skor akhir yang sama, padahal kemampuan pemecahan masalah mencakup berbagai tahapan berpikir. Akibatnya, guru belum memperoleh informasi yang memadai mengenai tahapan yang menjadi kesulitan utama siswa dalam menyelesaikan masalah matematika.

Sebagai alternatif, Modern Test Theory (MTT) atau Item Response Theory (IRT) menawarkan paradigma pengukuran yang lebih objektif melalui prinsip invariansi parameter. Pendekatan ini memungkinkan estimasi kemampuan peserta dan parameter butir diperoleh secara independen sehingga hasil pengukuran lebih stabil pada kelompok berbeda. Model Rasch memungkinkan transformasi data ordinal menjadi ukuran linear berbasis logit sehingga interpretasi kemampuan menjadi lebih objektif dan presisi (Ekstrand et al., 2022). Selain itu, model Rasch juga mampu mengevaluasi kualitas butir, reliabilitas instrumen, dan pola respons peserta secara lebih komprehensif dibandingkan pendekatan pengukuran klasik (Hagquist & Andrich, 2017).

Salah satu model Rasch yang sesuai untuk menganalisis instrumen uraian berskor bertingkat adalah Partial Credit Model (PCM). Model ini dirancang untuk mengakomodasi data politomus melalui

pemberian nilai berdasarkan kualitas tahapan penyelesaian siswa. PCM memungkinkan setiap langkah penyelesaian diperlakukan sebagai *item step* dengan tingkat kesulitan berbeda sehingga mampu mengidentifikasi rentang antar kategori skor secara lebih rinci (Bond et al., 2021). Penggunaan PCM relevan dalam pengukuran KPPM karena siswa dinilai tidak hanya berdasarkan jawaban akhir, tetapi juga kualitas proses berpikir selama penyelesaian soal.

Dalam penelitian ini, KPPM dianalisis berdasarkan tahapan pemecahan masalah Polya, yaitu memahami masalah, menyusun rencana, melaksanakan rencana, dan memeriksa kembali solusi. Menurut (Santos-Trigo, 2020), pendekatan Polya tetap relevan dalam pembelajaran matematika modern karena mampu menggambarkan struktur berpikir siswa dalam menyelesaikan masalah non-rutin. Integrasi tahapan Polya dengan PCM diharapkan menghasilkan instrumen pengukuran yang lebih komprehensif dalam mengevaluasi KPPM siswa.

Berbagai penelitian menunjukkan bahwa model Rasch politomus mampu meningkatkan kualitas evaluasi instrumen dibandingkan pendekatan klasik. Boone, (2016) menyatakan bahwa analisis Rasch efektif untuk mengevaluasi kualitas kategori skor dan stabilitas parameter butir secara objektif, sedangkan (Wahyuningsih, 2021) menunjukkan bahwa PCM sesuai digunakan pada instrumen uraian karena mampu mendeteksi fungsi kategori skor secara lebih rinci. Namun, implementasi PCM pada instrumen KPPM siswa SMP, khususnya materi Sistem Persamaan Linear Dua Variabel (SPLDV), masih relatif terbatas.

Sebagian besar penelitian sebelumnya lebih berfokus pada pengembangan instrumen atau analisis hasil belajar

matematika secara umum tanpa mengintegrasikan indikator pemecahan masalah Polya ke dalam evaluasi psikometrik modern secara komprehensif. Selain itu, kajian mengenai kualitas kategori skor politomus, parameter tingkat kesulitan butir, fungsi informasi instrumen, dan analisis Differential Item Functioning (DIF) masih jarang dilakukan. Padahal, menurut (Rupp, 2021), analisis DIF penting untuk memastikan bahwa keberhasilan peserta ditentukan oleh kemampuan laten, bukan faktor kelompok tertentu. Berdasarkan kondisi tersebut, penelitian ini berfokus pada evaluasi psikometrik instrumen KPPM berbasis tahapan Polya menggunakan PCM pada materi SPLDV SMP. Penelitian ini diharapkan memberikan kontribusi teoretis dalam pengembangan evaluasi matematika berbasis psikometrik modern sekaligus membantu guru menyusun sistem penilaian yang lebih objektif, akurat, dan diagnostik sesuai tuntutan pembelajaran abad ke-21.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan jenis penelitian deskriptif evaluatif yang bertujuan mengevaluasi kualitas psikometrik instrumen KPPM siswa menggunakan PCM. Penelitian dilaksanakan pada siswa SMP yang telah mempelajari materi Sistem Persamaan Linear Dua Variabel (SPLDV) dengan jumlah subjek sebanyak 176 siswa yang dipilih menggunakan teknik purposive sampling.

Instrumen penelitian berupa tes uraian berbasis tahapan pemecahan masalah menurut Polya yang meliputi memahami masalah, menyusun rencana, melaksanakan rencana, dan memeriksa kembali solusi. Instrumen dikembangkan berdasarkan kisi-kisi KPPM dan menggunakan rubrik penskoran politomus dengan rentang skor 0–3 untuk setiap

indikator. Pengumpulan data dilakukan melalui tes tertulis dan validasi ahli.

Data empiris dianalisis menggunakan model PCM, melalui bantuan perangkat lunak Winsteps (Kadek & Sumandya, 2019; Sumintono & Widhiarso, 2015).

Tahapan analisis meliputi uji prasyarat analisis, analisis kecocokan model, serta analisis reliabilitas person dan item (Bond et al., 2021; Dwinata, 2019). Selain itu, dilakukan analisis validitas empiris melalui point measure correlation, evaluasi fungsi kategori skor menggunakan *Andrich Threshold*, analisis tingkat kesulitan butir berdasarkan nilai item measure, interpretasi *Item Characteristic Curve* (ICC), serta analisis *Differential Item Functioning* (DIF) untuk mendeteksi kemungkinan bias butir berdasarkan jenis kelamin siswa (Helmiyatun et al., 2025).

HASIL DAN PEMBAHASAN

Hasil Validasi Ahli

Validasi instrumen dilakukan melalui penilaian tiga validator (dosen dan guru) untuk memastikan setiap butir soal sesuai dengan konstruk KPPM, kemudian dianalisis menggunakan indeks validitas isi Aiken's V guna memperoleh tingkat kesepakatan antarpenilai secara kuantitatif.

Tabel 1. Rangkuman Validasi Ahli

Butir Indikator	Ahli/Pakar			Aiken V	Kategori
	1	2	3		
1	5	5	5	1	Valid
2	5	5	4	0,92	Valid
3	5	5	5	1	Valid
4	5	5	5	1	Valid

Berdasarkan hasil analisis, seluruh butir memperoleh nilai Aiken's V yang lebih besar atau sama dengan 0,92. Menurut (Nurjanah et al., 2023), nilai ambang batas untuk menyatakan sebuah item adalah $V \geq 0,92$ berdasarkan kriteria

3 rater dan 5 kategori. Validitas isi instrumen ini telah memenuhi standar sebagai prasyarat awal sebelum dilakukan analisis empiris menggunakan Partial Credit Model (PCM).

Uji Prasyarat Analisis

Sebelum dilakukan estimasi parameter menggunakan Partial Credit Model (PCM), terlebih dahulu dilakukan pengujian asumsi model Rasch melalui analisis *Principal Component Analysis of Residuals* (PCAR). Pengujian ini bertujuan untuk memastikan bahwa instrumen hanya mengukur satu konstruk dominan, yaitu KPPM.

Tabel 2. *Principal Componen Analysis Residual*

Table of Standarized Residual Variance			
	Empirical		Modeled
Total raw variance in observations	9.7	100.0%	100.0%
by measures	5.7	58.8%	58.6%
by persons	2.9	29.5%	29.5%
by items	2.8	29.2 %	29.2%
Raw unexplained variance (total)	4.0	41.2%	41.4%
variance in 1st contrast	1.8	18.1%	

Hasil analisis yang dapat dilihat di Tabel 1 menunjukkan bahwa nilai *raw variance explained by measures* mencapai 58,8%. Nilai tersebut berada di atas batas minimum 40% sehingga menunjukkan bahwa instrumen memiliki konstruk utama yang kuat dan memenuhi asumsi unidimensionalitas (Linacre, 2021). Selain itu, nilai *unexplained variance in 1st contrast* sebesar 1,8 eigenvalue dengan kontribusi varians sebesar 18,1% mengindikasikan tidak adanya dimensi sekunder yang dominan dalam instrumen (Güler et al., 2021).

Dengan demikian, instrumen memenuhi asumsi unidimensi yang menunjukkan bahwa semua butir mengukur satu variabel kemampuan yang sama. Terpenuhinya

asumsi ini juga mengindikasikan bahwa independensi lokal telah tercapai, karena respons siswa terhadap suatu butir tidak dipengaruhi oleh respons pada butir lainnya (Debelak, 2019). Kondisi tersebut mendukung tercapainya invariansi parameter dalam model Rasch, sehingga estimasi tingkat kesulitan butir dan kemampuan siswa menjadi lebih objektif serta tidak bergantung pada karakteristik sampel tertentu (Debelak, 2019).

Uji Kecocokan Model (Model Fit)

Dalam menentukan kelayakan butir soal (*item fit*) instrumen didasarkan pada

kriteria statistik utama. Berdasarkan literatur pengukuran mutakhir (Bond et al., 2021; Linacre, 2025; Wind & Hua, 2022), persyaratan kecocokan model tersebut meliputi:

1. Nilai *Outfit Mean Square* (MNSQ) pada rentang 0,5 hingga 1,5
2. Nilai *Outfit Z-Standardized* (ZSTD) dengan batas toleransi -2,0 hingga +2,0.
3. Nilai *Point Measure Correlation* (Pt-Measure Corr) berada di antara 0,40 hingga 0,85.

Tabel 3. *Item Fit Order*

Item Number	Total Score	Measure	Model S.E.	Outfit		Pt-Measure	
				MNSQ	ZSTD	Corr.	Exp.
4	154	1.51	0.12	1.08	0.6	0.69	0.71
1	304	-0.30	0.11	1.1	0.9	0.65	0.71
2	332	-0.63	0.11	0.91	-0.8	0.73	0.70
3	327	-0.58	0.11	0.81	-1.8	0.75	0.70
MEAN	279.3	0	0.11	0.98	-0.3	-	-
S.D.	73.1	0.88	0	0.12	1.1	-	-

Berdasarkan hasil analisis data pada Tabel 3, seluruh butir soal menunjukkan nilai MNSQ, ZSTD, dan Pt-Measure Corr yang sepenuhnya berada di dalam rentang toleransi tersebut. Dengan demikian, dapat disimpulkan bahwa keseluruhan butir soal yang diujikan memenuhi syarat kecocokan model (*model fit*) secara komprehensif, sehingga instrumen ini terbukti layak dan data yang telah diambil dapat dilanjutkan ke tahap analisis selanjutnya.

Uji Reliabilitas

Analisis reliabilitas dilakukan untuk mengevaluasi konsistensi instrumen dalam mengukur KPPM. Nilai *Item Reliability* dan *person reliability*

Tabel 4. *Reliabilitas Person dan Item*

Aspek	RMSE	True SD	Separation	Reliability
Person	0.8	1.3	1.62	0.72
Item	0.11	0.87	7.78	0.98

Berdasarkan hasil analisis menggunakan Winsteps, diperoleh nilai *person reliability* sebesar 0,72 dan *item reliability* sebesar 0,98. Menurut kriteria reliabilitas dalam model Rasch, nilai *person reliability* termasuk kategori cukup, sedangkan nilai *item reliability* berada pada kategori istimewa atau *excellent* (Sumintono & Widhiarso, 2015). Dengan demikian, instrumen yang digunakan dapat dinyatakan memiliki kualitas butir yang sangat baik dan layak digunakan sebagai alat ukur penelitian.

Validitas Instrumen Berdasarkan Data

Validitas empiris instrumen dianalisis melalui statistik *item fit order* dan korelasi *Point-Measure* untuk memastikan bahwa setiap butir benar-benar mengukur konstruk KPPM. Hasil analisis yang ada

pada Tabel 3 menunjukkan bahwa seluruh butir memenuhi syarat yang diberikan. Validitas empiris juga diperkuat melalui analisis kategori skor politomus yang dapat dilihat dari nilai *Observed Sample Average* yang meningkat dari kategori skor 0 hingga 3.

Tabel 5. Fungsi Kategori Skor

Category Label	Observed Sample Average	Observed Sample Expect	Infit Mnsq	Outfit Mnsq	Andrich Threshold	Category Measure
0	-1.87	-1.85	0.98	0.98	NONE	-2.61
1	-0.55	-0.54	0.97	0.97	-1.32	-0.86
2	0.69	0.62	0.86	0.84	-0.24	0.77
3	1.72	1.78	1.14	1.12	1.55	2.76

Selain itu, parameter *Andrich Threshold* menunjukkan pola peningkatan yang teratur tanpa adanya *disordered threshold*, yang mengindikasikan bahwa kategori skor telah berfungsi secara logis dan sesuai dengan perkembangan kemampuan responden (Bond et al., 2021).

Nilai *Category Measure* bergerak meningkat secara monoton dari -2.61 logit pada kategori 0, naik menjadi -0.86 logit pada kategori 1, 0.77 logit pada kategori 2, hingga mencapai 2.76 logit pada kategori 3, menegaskan bahwa setiap kenaikan kategori skor merepresentasikan peningkatan kemampuan yang konsisten pada responden (Bond et al., 2021).

Karakteristik Parameter Butir Soal

Analisis karakteristik parameter butir merupakan bagian penting dalam evaluasi psikometrik berbasis *Item Response Theory* (IRT), khususnya pada penggunaan *Partial Credit Model* (PCM). Analisis ini bertujuan untuk mengidentifikasi bagaimana setiap indikator dalam instrumen berfungsi dalam mengukur kemampuan laten siswa secara objektif dan konsisten. Berbeda dengan pendekatan pengukuran klasik yang hanya menitikberatkan pada skor

total, PCM memungkinkan peneliti mengevaluasi kualitas setiap tahapan respons siswa melalui estimasi parameter logit, kurva karakteristik butir (*Item Characteristic Curve/ICC*), serta pengujian bias butir (*Differential Item Functioning/DIF*) (Aryadoust et al., 2021; Bond et al., 2021).

1. Parameter Tingkat Kesulitan Indikator (soal)

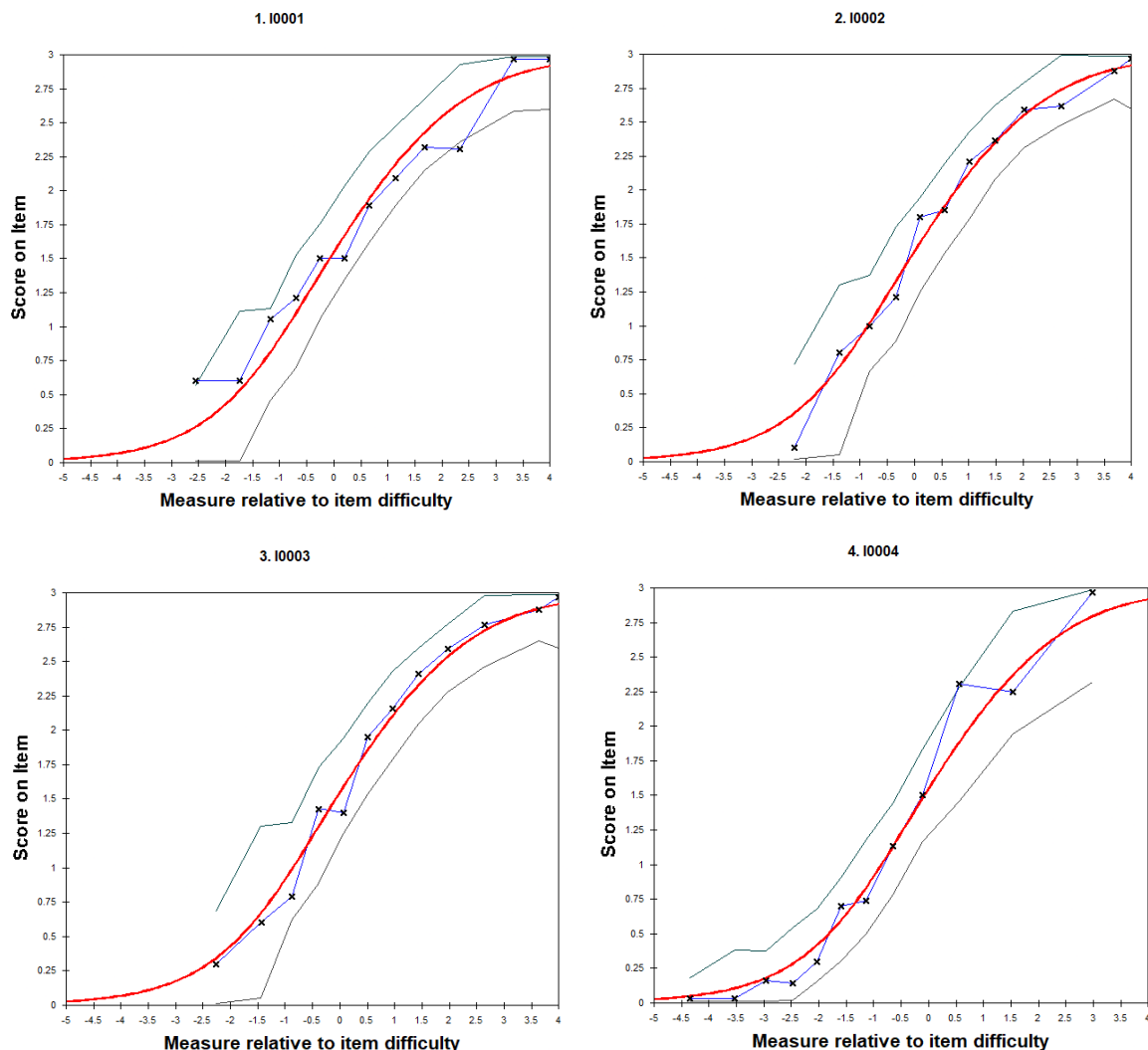
Parameter tingkat kesulitan dalam PCM menunjukkan posisi relatif suatu indikator pada skala logit. Semakin tinggi nilai logit suatu butir, maka semakin tinggi pula kemampuan yang diperlukan siswa untuk memperoleh skor maksimal pada indikator tersebut (Bond et al., 2021). Berdasarkan hasil estimasi *item measure* menggunakan aplikasi Winsteps, diperoleh bahwa keempat indikator kemampuan pemecahan masalah matematis memiliki tingkat kesulitan yang bervariasi, yaitu berkisar antara -0,64 logit hingga +1,51 logit. Variasi ini menunjukkan bahwa instrumen mampu memetakan kemampuan siswa secara bertingkat dari level dasar hingga level berpikir tingkat tinggi. Kondisi ini sejalan dengan karakteristik PCM yang mampu mengevaluasi tahapan respons peserta didik secara lebih rinci melalui

estimasi parameter logit pada setiap kategori skor (Wahyuningsih, 2021).

2. Item Characteristic Curve (ICC)

Analisis *Item Characteristic Curve* (ICC) dilakukan untuk melihat kesesuaian antara data empiris siswa dengan prediksi model Rasch pada setiap indikator kemampuan pemecahan masalah matematis. Kurva

ICC menggambarkan hubungan antara kemampuan siswa (θ) dengan probabilitas memperoleh skor tertentu pada suatu indikator. Dalam model PCM, kurva ini menjadi indikator penting untuk mengevaluasi apakah suatu butir bekerja secara konsisten dan objektif dalam mengukur kemampuan siswa (Bond et al., 2021).



Gambar 1. Grafik *Item Characteristic Curve* tiap butir item

Berdasarkan hasil analisis grafik ICC, seluruh indikator menunjukkan pola kurva yang mengikuti model teoritis secara konsisten. Sebaran titik observasi empiris berada di sekitar garis estimasi model dan tetap berada dalam interval kepercayaan 95%, sehingga mengindikasikan bahwa

seluruh indikator memiliki tingkat kecocokan (*fit*) yang baik terhadap model PCM. Kondisi ini menunjukkan bahwa probabilitas siswa memperoleh skor yang lebih tinggi meningkat secara monoton seiring meningkatnya kemampuan laten yang dimiliki (Aryadoust et al., 2021).

Secara keseluruhan, konsistensi pola seluruh grafik ICC membuktikan bahwa instrumen berbasis langkah Polya yang dikembangkan telah terkalibrasi dengan baik secara empiris. Dengan demikian, setiap skor yang diperoleh siswa dapat diinterpretasikan sebagai representasi yang valid terhadap kemampuan pemecahan masalah matematis yang dimiliki (Boone et al., 2016).

3. Differential Item Functioning (DIF)

Tabel 6. Summary DIF Mantel-Haenszel based on Gender

Item	Laki-laki		Perempuan		DIF CONTRAST	Mantel-Haenszel Chi-squ	Prob.
	DIF MEASURE	DIF S.E.	DIF MEASURE	DIF S.E.			
I0001	-0.15	0.16	-0.41	0.15	0.26	1.6124	2,042
I0002	-0.64	0.16	-0.64	0.15	0.00	3	9,872
I0003	-0.67	0.16	-0.51	0.15	-0.17	6,401	4,237
I0004	1.47	0.17	1.57	0.16	-0.11	3,102	5,776

Interpretasi hasil *Differential Item Functioning* (DIF) menunjukkan bahwa tidak terdapat perbedaan fungsi butir yang signifikan antara kelompok laki-laki (L) dan perempuan (P) pada keempat item yang dianalisis. Dalam analisis DIF berbasis model Rasch, suatu item dinyatakan mengalami DIF apabila nilai signifikansi (*p-value*) lebih kecil dari 0,05, sedangkan nilai $p > 0,05$ menunjukkan bahwa perbedaan fungsi item antar kelompok tidak signifikan secara statistik (Hagquist & Andrich, 2017).

Baik pada item I0001, I0003, dan I0004 menunjukkan perbedaan yang dapat dinyatakan tidak signifikan secara statistik berdasarkan kriteria yang di gunakan. Sementara itu, pada item I0002, nilai DIF yang diperoleh identik, yaitu -0,63 pada kedua kelompok, sehingga menghasilkan kontras DIF sebesar nol dengan nilai $t = 0$ dan $p = 1,00$. Hasil ini menunjukkan bahwa item tersebut berfungsi sama pada kedua kelompok dan tidak mengandung bias gender.

seluruh item memiliki nilai probabilitas (*p-value*) lebih besar dari 0,05 sehingga tidak terdapat bukti yang menyatakan adanya

Analisis *Differential Item Functioning* (DIF) dilakukan untuk menguji apakah setiap butir soal bekerja secara adil terhadap kelompok responden yang berbeda. Dalam penelitian ini, pengujian DIF dilakukan berdasarkan gender untuk memastikan bahwa probabilitas siswa memperoleh skor tertentu hanya dipengaruhi oleh kemampuan matematisnya, bukan oleh faktor keanggotaan kelompok.

bias butir berdasarkan gender. Dalam analisis Rasch, kondisi ini menunjukkan bahwa item tidak mengalami *Differential Item Functioning* (DIF) yang signifikan dan mampu mengukur konstruk secara setara pada kelompok yang berbeda (Kleinman et al., 2017). Dengan demikian, keempat item dapat dinyatakan memenuhi prinsip invariansi pengukuran dan dapat digunakan secara adil pada kelompok laki-laki maupun perempuan.

KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa instrumen kemampuan pemecahan masalah matematis (KPPM) berbasis tahapan Polya pada materi Sistem Persamaan Linear Dua Variabel (SPLDV) telah memenuhi kriteria psikometrik modern melalui analisis Partial Credit Model (PCM). Hasil validasi ahli menunjukkan seluruh butir memiliki validitas isi yang sangat tinggi, sedangkan hasil uji unidimensi membuktikan bahwa instrumen mengukur satu konstruk kemampuan yang sama secara konsisten. Analisis kecocokan model menunjukkan seluruh item memenuhi kriteria fit

berdasarkan nilai Outfit MNSQ, ZSTD, dan Point Measure Correlation, sehingga setiap indikator dinyatakan bekerja sesuai dengan model Rasch. Dari aspek reliabilitas, instrumen memiliki kualitas yang baik dengan nilai person reliability sebesar 0,72 dan item reliability sebesar 0,98, yang menandakan konsistensi pengukuran dan stabilitas parameter butir sangat tinggi. Analisis kategori skor politomus memperlihatkan bahwa rubrik penskoran mampu membedakan tingkat kemampuan siswa secara logis dan bertahap tanpa terjadi disordered threshold. Variasi parameter tingkat kesulitan butir menunjukkan bahwa instrumen mampu mengukur kemampuan siswa dari tingkat rendah hingga tinggi secara objektif. Selain itu, hasil Item Characteristic Curve (ICC) menunjukkan seluruh item mengikuti pola teoritis model PCM dan mampu merepresentasikan hubungan antara kemampuan siswa dengan probabilitas memperoleh skor secara konsisten. Hasil analisis Differential Item Functioning (DIF) juga membuktikan bahwa seluruh butir bebas dari bias gender sehingga memenuhi prinsip keadilan pengukuran. Dengan demikian, instrumen KPPM berbasis PCM yang dikembangkan dinyatakan valid, reliabel, objektif, dan layak digunakan sebagai alat evaluasi kemampuan pemecahan masalah matematis siswa SMP secara lebih diagnostik dan komprehensif dibandingkan pendekatan pengukuran klasik.

DAFTAR PUSTAKA

- Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A Comprehensive Review Of Rasch Measurement In Language Assessment: Recommendations And Guidelines For Research. *Language Testing*, 38(1), 6–40. <https://doi.org/10.1177/0265532220927487>
- Bond, T., Yan, Z., & Heene, M. (2021). *Applying The Rasch Model: Fundamental Measurement In The Human Sciences* (4th Ed.). Routledge.
- Boone, W. J. (2016). Rasch Analysis For Instrument Development: Why, When, And How? *CBE Life Sciences Education*, 15(4). <https://doi.org/10.1187/Cbe.16-04-0148>
- Boone, W. J., Staver, J. R., & Yale, M. S. (2016). *Rasch Analysis In The Human Sciences*. Springer. <https://doi.org/10.1007/978-94-007-6857-4>
- Debelak, R. (2019). An Evaluation Of Overall Goodness-Of-Fit Tests For The Rasch Model. *Frontiers In Psychology*, 9(JAN). <https://doi.org/10.3389/fpsyg.2018.02710>
- Dwinata, A. (2019). Analisis Kemampuan Pemecahan Masalah Menggunakan Pemodelan Rasch Pada Materi Permutasi Dan Kombinasi. *PRISMA, Prosiding Seminar Nasional Matematika*, 2, 124–131.
- Ekstrand, J., Westergren, A., Årestedt, K., Hellström, A., & Hagell, P. (2022). Transformation Of Rasch Model Logits For Enhanced Interpretability. *BMC Medical Research Methodology*, 22(1), 1–10. <https://doi.org/10.1186/S12874-022-01816-1>
- Faradila, O. Vini, Falani, I., & Mujahidawati, M. (2025). Psychometric Evaluation Of A Metaphorical Thinking Instrument In Mathematics Learning: Graded Response Model. *Journal Of Mathematics Instruction, Social Research And Opinion*, 4(4), 1453–1464. <https://doi.org/10.58421/Misro.V4i4.873>

- Güler, N., Uyanık, G. K., & Teker, G. T. (2021). Validity And Reliability Study Using Rasch Measurement Model. *Education And Information Technologies*, 26, 3105–3121. <https://doi.org/10.1007/S10639-020-10428-7>
- Hagquist, C., & Andrich, D. (2017). Recent Advances In Analysis Of Differential Item Functioning In Health Research Using The Rasch Model. *Health And Quality Of Life Outcomes*, 15(1), 1–8. <https://doi.org/10.1186/S12955-017-0755-0>
- Helmiyatun, Taqiudin, M., & Ahmad. (2025). Content Validity Of Student Perception Instrument On School Educational Management. *Journal Of Practice Learning And Educational Development*, 5(1), 93–103. <https://doi.org/10.58737/Jpled.V5i1.422>
- Kadek, N. R. P., & Sumandya, I. W. (2019). Penerapan Partial Credit Model (PCM) Dalam Mengevaluasi Tes Uraian Application Of Partial Credit Model (PCM) In Evaluating The Essay Tes. *Jurnal EMASAINS Volume III, VIII*, 77–85.
- Kleinman, M., Teresi, J. A., Health, R., & Cornell, W. (2017). *From Item Response Theory Models*. 58(1), 79–98.
- Linacre, J. M. (2021). *Principal Components Analysis Of Residuals*. Winsteps Rasch Measurement. <https://www.winsteps.com/winman/principalcomponents.htm>
- Linacre, J. M. (2025). *A User's Guide To WINSTEPS Ministep: Rasch-Model Computer Programs*. Winsteps.Com.
- Nurjanah, S., Istiyono, E., Widiastuti, W., Iqbal, M., & Kamal, S. (2023). The Application Of Aiken's V Method For Evaluating The Content Validity Of Instruments That Measure The Implementation Of Formative Assessments. *Journal Of Research And Educational Research Evaluation*, 12(2), 2023–2125. <http://journal.unnes.ac.id/sju/index.php/jere>
- Rupp, A. A. (2021). ITEMS Corner ITEMS Portal Development Update <https://ncme.elevate.com/partners.com>. *Educational Measurement: Issues And Practice*, 40(2), 106–107. <https://doi.org/10.1111/emip.12436>
- Santos-Trigo, M. (2020). Problem-Solving In Mathematics Education. In *Encyclopedia Of Mathematics Education* (Pp. 686–693). Springer.
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch Pada Asesmen Pendidikan*. Trim Komunikata.
- Wahyuningsih, S. (2021). Using The Rasch's Partial Credit Model To Analyze The Quality Of An Essay Math Test. *Proceedings Of The 1st International Conference On Mathematics And Mathematics Education (Icmmmed 2020)*, 550(Icmmmed 2020), 257–265. <https://doi.org/10.2991/assehr.k.210508.073>
- Wind, S. A., & Hua, C. (2022). *Rasch Measurement Theory Analysis In R* (1st Ed.). Chapman And Hall/CRC.