

ANALISIS KUALITAS INSTRUMEN ASESMEN SAINS SISWA KELAS V SEKOLAH DASAR DALAM KONTEKS *DEEP LEARNING* MENGGUNAKAN MODEL RASCH

Dewi Juita¹, Zikra², Putri Zuhra³, Yusmaridi M⁴, Festiyed⁵, Skunda Diliarosta⁶, Fatni Mufit⁷

Universitas Islam Negeri Imam Bonjol Padang¹, Universitas PGRI Sumatera Barat^{1,2}, Universitas Negeri Padang^{2,3}, Sekolah Dasar Negeri Percobaan Padang⁴, Universitas Negeri Padang^{5,6,7}
dewijuita@uinib.ac.id¹

ABSTRAK

Penelitian ini bertujuan untuk mengevaluasi kualitas psikometrik instrumen asesmen IPA bagi siswa kelas V sekolah dasar dalam konteks pembelajaran *deep learning* menggunakan model Rasch. Penelitian ini menggunakan pendekatan deskriptif kuantitatif dengan analisis Rasch satu parameter logistik (1-PL) menggunakan perangkat lunak Winsteps versi 5.10.4.0 terhadap 20 butir soal pilihan ganda dikotomis yang diujikan kepada 29 siswa kelas V sekolah dasar; analisis mencakup tingkat kesukaran butir, validitas konstruk, unidimensionalitas berbasis PCA *residual*, kemampuan responden, kesesuaian responden (*person fit*), dan reliabilitas instrumen. Hasil penelitian menunjukkan reliabilitas yang tinggi dengan koefisien KR-20 sebesar 0,89, reliabilitas responden 0,84, dan reliabilitas butir 0,81; varians mentah yang dijelaskan oleh ukuran mencapai 42,9% melampaui standar Rasch 40% dan seluruh kontras varians tak terjelaskan berada di bawah 10% sehingga instrumen terbukti unidimensional; sebanyak 19 dari 20 butir dinyatakan valid sedangkan butir A10 mengalami *misfit* (*Outfit* MNSQ = 2,45; ZSTD = 2,15); serta distribusi kemampuan siswa menunjukkan 20,69% berkategori tinggi, 62,07% sedang, dan 17,24% rendah dengan enam siswa menunjukkan pola respons tidak konsisten. Penelitian ini menyimpulkan bahwa instrumen asesmen IPA yang dikembangkan dalam kerangka *deep learning* memiliki kualitas psikometrik yang memadai untuk mengukur kompetensi IPA siswa kelas V sekolah dasar.

Kata Kunci: Asesmen IPA, *Deep Learning*, Model Rasch, Teori Respons Butir, Unidimensionalitas, Validitas Konstruk.

ABSTRACT

This study aimed to evaluate the psychometric quality of a science assessment instrument designed for fifth-grade elementary school students in a deep learning instructional context using the Rasch model. This study used a descriptive quantitative approach with a one-parameter logistic Rasch model (1-PL) analyzed using Winsteps software version 5.10.4.0 applied to 20 dichotomous multiple-choice items administered to 29 fifth-grade elementary school students; the analysis covered item difficulty, construct validity, unidimensionality based on PCA of standardized residuals, person ability, person fit, and

instrument reliability. The results showed strong reliability with a KR-20 coefficient of 0.89, person reliability of 0.84, and item reliability of 0.81; the raw variance explained by the measures reached 42.9% exceeding the 40% Rasch standard and all unexplained variance contrasts remained below 10% confirming that the instrument is unidimensional; 19 of the 20 items were valid while item A10 showed misfit (Outfit MNSQ = 2.45; ZSTD = 2.15); and the distribution of student ability indicated 20.69% in the high category, 62.07% moderate, and 17.24% low with six students exhibiting inconsistent response patterns. This study concludes that the science assessment instrument developed within the deep learning framework demonstrates adequate psychometric quality for measuring fifth-grade elementary school students' science competence.

Keywords: *Deep Learning, Item Response Theory, Rasch Model, Science Assessment, Unidimensionality, Construct Validity.*

PENDAHULUAN

Kualitas instrumen asesmen pendidikan memegang peranan penting dalam menjamin validitas keputusan pedagogis yang didasarkan pada hasil pengukuran (Cook & Beckman, 2006). Dalam pendidikan IPA di tingkat sekolah dasar, isu ini menjadi semakin penting karena pengukuran yang akurat terhadap kompetensi sains siswa menjadi landasan dalam merancang pembelajaran yang tepat sasaran, mengevaluasi kurikulum, serta meneliti perkembangan belajar siswa (Oliva & Blanco-Lopez, 2023). Instrumen yang tidak memenuhi kriteria psikometrik yang ditetapkan dapat menyebabkan estimasi kemampuan siswa yang tidak akurat, menurunkan keterinterpretasian skor, serta berpotensi menghasilkan praktik pendidikan yang kurang tepat (Bond & Fox, 2015).

Konsep *deep learning* yang didefinisikan sebagai pendekatan pembelajaran yang menekankan pemahaman konseptual, pemikiran reflektif, dan penerapan pengetahuan dalam konteks baru semakin memperoleh perhatian dalam reformasi pendidikan mutakhir, termasuk dalam Kurikulum Merdeka di Indonesia (Fullan et al., 2018; Quinn et al., 2019). Berbeda dengan *surface learning* yang bertumpu pada hafalan, *deep learning* menuntut peserta didik untuk secara aktif membangun makna, terlibat dalam berpikir tingkat tinggi, serta menghubungkan pengetahuan baru dengan pengalaman sebelumnya (Asikainen & Gijbels, 2017; Winje & Londal, 2023). Bukti empiris menunjukkan adanya hubungan timbal balik antara desain asesmen dan pendekatan belajar: asesmen yang menekankan pemahaman cenderung mendorong strategi belajar yang lebih mendalam, sedangkan asesmen yang berfokus pada hafalan justru mendorong keterlibatan belajar pada tingkat permukaan (Lau & Liem, 2024; Postareff et al., 2019). Keterkaitan ini menegaskan pentingnya memastikan bahwa instrumen asesmen dalam lingkungan *deep learning* tidak hanya relevan secara konten, tetapi juga memiliki kualitas psikometrik yang memadai.

Metode evaluasi tradisional dalam *Classical Test Theory* (CTT) memiliki keterbatasan karena sangat bergantung pada sampel tertentu sehingga indeks kesukaran dan daya pembeda butir dapat berubah sesuai karakteristik peserta tes dan tidak dapat

digeneralisasikan ke berbagai konteks (Hambleton et al., 2016). Sebaliknya, model Rasch yang merupakan bentuk paling parsimonious dalam *Item Response Theory* (IRT) mampu mengatasi keterbatasan tersebut dengan memetakan tingkat kesukaran butir dan kemampuan responden ke dalam satu metrik skala interval yang sama (logit) dengan estimasi yang secara teoretis bersifat invarian antar populasi (Boone, 2016; Tesio et al., 2023). Kekuatan penting lain dari model Rasch adalah kemampuannya menilai unidimensionalitas melalui *Principal Component Analysis* (PCA) of standardized residuals, yang digunakan untuk mengevaluasi apakah varians yang tersisa setelah konstruk laten utama diekstraksi cukup kecil untuk mendukung kesimpulan bahwa instrumen mengukur satu konstruk tunggal (Reeve et al., 2007; Smith, 2002).

Meskipun memiliki keunggulan metodologis tersebut, penerapan analisis Rasch untuk memvalidasi instrumen asesmen IPA di tingkat sekolah dasar di Indonesia masih relatif terbatas. Studi-studi yang ada tentang HOTS di sekolah dasar telah menunjukkan pentingnya pengembangan instrumen berbasis kemampuan berpikir tingkat tinggi sejak jenjang SD (Fanani & Kusmaharti, 2018; Isrokatun et al., 2024; Purwaningsih et al., 2022), namun validasi psikometriknya umumnya masih bertumpu pada pendekatan CTT dan belum mengintegrasikan kerangka *deep learning* sebagai landasan konseptual. Penelitian Darmana et al. (2024) yang memvalidasi instrumen sains berbasis Rasch dilakukan pada siswa SMA, sedangkan Yudha (2023) menerapkan analisis Rasch pada jenjang SMP — keduanya tidak menyentuh jenjang sekolah dasar yang memiliki karakteristik perkembangan kognitif berbeda secara fundamental. Selain itu, sebagian besar studi terdahulu hanya melaporkan sebagian parameter *fit* tanpa menyajikan analisis unidimensionalitas berbasis PCA *residual*, *person fit*, dan reliabilitas butir dalam satu kerangka pelaporan yang terpadu (Hamzah et al., 2022; Scoulas et al., 2021). Namun, penelitian sebelumnya belum membahas secara terpadu pengembangan dan validasi instrumen asesmen IPA berbasis *Higher Order Thinking Skills* (HOTS) dalam kerangka *deep learning* untuk siswa kelas V sekolah dasar menggunakan analisis Rasch yang komprehensif. Kebaruan penelitian ini adalah integrasi kerangka *deep learning* sebagai basis pengembangan instrumen asesmen IPA untuk jenjang sekolah dasar disertai penerapan analisis Rasch secara menyeluruh mencakup tingkat kesukaran butir, validitas konstruk, unidimensionalitas berbasis PCA *residual*, *person fit*, dan reliabilitas dalam satu studi terpadu yang belum pernah dilaporkan pada konteks dan jenjang yang sama di Indonesia. Penelitian ini bertujuan untuk mengevaluasi karakteristik psikometrik instrumen asesmen IPA ditinjau dari tingkat kesukaran butir, validitas konstruk, unidimensionalitas, distribusi kemampuan siswa, kesesuaian responden, dan reliabilitas menggunakan model Rasch.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan deskriptif kuantitatif untuk mengkaji karakteristik psikometrik instrumen asesmen IPA. Instrumen yang digunakan terdiri atas 20 butir soal pilihan ganda dengan skor dikotomis (0/1) yang dikembangkan berdasarkan standar kurikulum IPA kelas V sekolah dasar dalam kerangka *deep learning* dengan

mengacu pada prinsip *Higher Order Thinking Skills* (HOTS), sehingga butir yang dikembangkan diarahkan untuk mengukur kemampuan pemahaman, penerapan, dan penalaran analitis, bukan sekadar kemampuan mengingat (Yudha, 2023). Penelitian dilaksanakan di (*mohon dilengkapi: nama sekolah dan lokasi*) pada (*mohon dilengkapi: bulan dan tahun*). Populasi penelitian adalah seluruh siswa kelas V sekolah dasar tersebut. Sampel penelitian berjumlah 29 siswa yang dipilih menggunakan teknik (*mohon dilengkapi: teknik sampling*). Meskipun ukuran sampel ini relatif kecil dibandingkan penelitian psikometrik berskala besar, penelitian sebelumnya menunjukkan bahwa estimasi model Rasch tetap dapat menghasilkan temuan yang stabil dan dapat diinterpretasikan pada sampel sekitar 25–30 responden, khususnya ketika menggunakan butir soal dikotomis (Darmana et al., 2024; Linacre, 2023).

Data respons siswa dianalisis menggunakan model Rasch satu parameter logistik (1-PL) dengan bantuan perangkat lunak Winsteps versi 5.10.4.0. Analisis dilakukan melalui enam tahapan: (1) kalibrasi tingkat kesukaran butir (*item difficulty*) yang dinyatakan dalam satuan logit pada skala yang sama dengan kemampuan responden; (2) evaluasi validitas konstruk melalui analisis polaritas butir (*Point Measure Correlation/Pt. Corr*) dan statistik *item fit*; (3) pengujian unidimensionalitas menggunakan *Principal Component Analysis* (PCA) of standardized residuals; (4) pengelompokan kemampuan responden berdasarkan kriteria *mean* $\pm$ standar deviasi; (5) pemeriksaan kesesuaian responden (*person fit*); dan (6) estimasi reliabilitas instrumen yang meliputi KR-20, reliabilitas responden (*person reliability*), dan reliabilitas butir (*item reliability*).

Kriteria penerimaan data mengacu pada pedoman Sumintono dan Widhiarso (2015), yaitu nilai *Outfit Mean Square* (MNSQ) pada rentang 0,5–1,5, nilai *Outfit Z-Standard* (ZSTD) pada rentang –2,0 hingga +2,0, dan nilai *Point Measure Correlation* (Pt. Corr) pada rentang 0,40–0,85. Suatu butir dinyatakan mengalami *misfit* apabila secara bersamaan melanggar kriteria *Outfit* MNSQ dan ZSTD. Untuk pengujian unidimensionalitas, kriteria yang digunakan meliputi persentase varians yang dijelaskan oleh ukuran minimal 20% dengan $\geq 40\%$ sebagai standar ideal, serta seluruh nilai *unexplained variance contrasts* berada di bawah 10% (Misbach & Sumintono, 2014; Wibisono, 2014). Interpretasi reliabilitas mengacu pada klasifikasi Sumintono dan Widhiarso (2015), di mana nilai KR-20 $> 0,80$ menunjukkan reliabilitas yang sangat baik.

HASIL PENELITIAN

Tingkat Kesukaran Butir

Dalam model Rasch, tingkat kesukaran butir dan kemampuan responden ditempatkan pada satu skala logit yang sama, dengan rata-rata tingkat kesukaran butir umumnya ditetapkan pada 0,00 logit (Boone, 2016). Berdasarkan skala tersebut, butir soal diklasifikasikan ke dalam tiga kategori, yaitu sukar (*Measure* $> +1,20$), sedang ($-1,20$ hingga $+1,20$), dan mudah (*Measure* $< -1,20$), menggunakan kriteria rata-rata $\pm$ satu standar deviasi ($SD = 1,20$). Pendekatan klasifikasi ini sejalan dengan praktik yang umum digunakan dalam pelaporan tingkat kesukaran butir berbasis Rasch (Hamzah et al., 2022; Yudha, 2023). Hasil analisis tingkat kesukaran butir disajikan pada Tabel 1.

Tabel 1. Tingkat Kesukaran dan Statistik Kesesuaian Butir (*Winsteps Table 13.1 & Table 10*)

No	Item	Raw Score	N	Measure (Logit)	S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. Corr	Kategori Kesukaran
1	A3	8	29	1.82	.51	.93	-.16	.80	-.04	.57	Sukar
2	A20	8	29	1.82	.51	.74	-.94	.54	-.51	.65	Sukar
3	A17	9	29	1.57	.49	1.05	.28	.97	.17	.53	Sukar
4	A10	11	29	1.11	.47	1.18	.80	2.45	2.15	.43	Sedang
5	A4	12	29	.89	.47	1.11	.56	.92	.00	.54	Sedang
6	A6	12	29	.89	.47	1.07	.39	.93	.03	.55	Sedang
7	A1	13	29	.67	.46	1.21	1.01	1.38	.89	.48	Sedang
8	A8	13	29	.67	.46	.72	-1.44	.52	-1.12	.72	Sedang
9	A9	14	29	.46	.46	1.14	.71	1.70	1.46	.49	Sedang
10	A11	16	29	.04	.46	1.27	1.29	1.41	.96	.46	Sedang
11	A15	16	29	.04	.46	.90	-.42	.66	-.71	.65	Sedang
12	A2	18	29	-.40	.48	.79	-.94	.63	-.66	.68	Sedang
13	A18	18	29	-.40	.48	.91	-.32	.77	-.32	.62	Sedang
14	A7	19	29	-.63	.49	1.13	.57	1.63	1.13	.49	Sedang
15	A14	19	29	-.63	.49	.94	-.17	.73	-.34	.62	Sedang
16	A5	21	29	-1.13	.52	1.08	.36	1.19	.49	.51	Mudah
17	A13	21	29	-1.13	.52	.98	.03	.76	-.14	.58	Mudah
18	A12	22	29	-1.41	.54	1.19	.67	.88	.12	.49	Mudah
19	A16	22	29	-1.41	.54	.61	-1.34	.42	-.65	.70	Mudah
20	A19	26	29	-2.85	.69	.68	-.77	.23	-.35	.56	Mudah
Mean		15.9	29	.00	.50	.98	.01	.98	.13		SD=1.20

Catatan. Baris berwarna merah = *misfit* (Outfit MNSQ > 1,5 dan ZSTD > 2,0 dilanggar secara bersamaan); SE = *Standard Error*; Pt. Corr = *Point-Measure Correlation*.

Berdasarkan Tabel 1, terdapat tiga butir (A3, A20, dan A17) yang tergolong sukar, dengan nilai logit antara 1.57 hingga 1.82. Hal ini menunjukkan bahwa ketiga butir tersebut menuntut tingkat penalaran ilmiah yang lebih tinggi dibandingkan kemampuan sebagian besar siswa dalam sampel. Sebanyak tiga belas butir berada pada kategori sedang, yang menunjukkan bahwa mayoritas soal telah sesuai dengan tingkat kemampuan target peserta didik. Sementara itu, empat butir tergolong mudah.

Rentang kesukaran butir membentang dari -2.85 hingga +1.82 logit. Butir A19 merupakan butir termudah (logit -2.85) dengan tingkat keberhasilan sekitar 90%, sementara butir A3 dan A20 merupakan butir tersukar (logit +1.82). Nilai standard error (SE) butir berada pada rentang 0.46 hingga 0.69 logit. Secara keseluruhan, distribusi tingkat kesukaran butir mencakup rentang yang memadai untuk merepresentasikan variasi kemampuan siswa dalam sampel.

Validitas Konstruk

Dalam kerangka pengukuran Rasch, validitas konstruk dapat ditinjau melalui dua indikator utama yang saling melengkapi. Indikator pertama adalah polaritas butir, yang diukur melalui Point-Measure Correlation (Pt. Corr). Indikator ini menunjukkan arah dan kekuatan kontribusi setiap butir terhadap konstruk laten yang diukur. Indikator kedua adalah statistik kesesuaian butir (*item fit statistics*), khususnya Outfit Mean Square (MNSQ) dan Outfit Z-Standard (ZSTD), yang menggambarkan sejauh mana pola respons aktual sesuai dengan ekspektasi model Rasch (Cook & Beckman, 2006; Reeve et al., 2007).

Seluruh butir diharapkan memiliki nilai Pt. Corr positif, karena nilai negatif menunjukkan bahwa suatu butir tidak bekerja secara konsisten dengan konstruk yang hendak diukur (Wolfe & Smith, 2007). Ringkasan hasil analisis validitas konstruk per butir disajikan pada Tabel 2.

Tabel 2. Ringkasan Validitas Konstruk per Butir (*Winsteps Table 10*)

Item	Pt. Corr (> 0)	Outfit MNSQ (0.5–1.5)	Outfit ZSTD (±2.0)	Kriteria Pt. Corr	Kriteria Fit	Kesimpulan
A1	.48	1.38	.89	Memenuhi	Memenuhi	Valid
A2	.68	.63	-.66	Memenuhi	Memenuhi	Valid
A3	.57	.80	-.04	Memenuhi	Memenuhi	Valid
A4	.54	.92	.00	Memenuhi	Memenuhi	Valid
A5	.51	1.19	.49	Memenuhi	Memenuhi	Valid
A6	.55	.93	.03	Memenuhi	Memenuhi	Valid
A7	.49	1.63	1.13	Memenuhi	Memenuhi	Valid
A8	.72	.52	-1.12	Memenuhi	Memenuhi	Valid
A9	.49	1.70	1.46	Memenuhi	Memenuhi	Valid
A10	.43	2.45	2.15	Memenuhi	Tidak Memenuhi	Tidak Valid
A11	.46	1.41	.96	Memenuhi	Memenuhi	Valid
A12	.49	.88	.12	Memenuhi	Memenuhi	Valid
A13	.58	.76	-.14	Memenuhi	Memenuhi	Valid
A14	.62	.73	-.34	Memenuhi	Memenuhi	Valid
A15	.65	.66	-.71	Memenuhi	Memenuhi	Valid
A16	.70	.42	-.65	Memenuhi	Memenuhi	Valid
A17	.53	.97	.17	Memenuhi	Memenuhi	Valid
A18	.62	.77	-.32	Memenuhi	Memenuhi	Valid
A19	.56	.23	-.35	Memenuhi	Memenuhi	Valid
A20	.65	.54	-.51	Memenuhi	Memenuhi	Valid

Catatan. Baris berwarna merah = butir tidak valid; kriteria Pt. Corr: positif pada rentang 0.40–0.85; kriteria Outfit MNSQ: 0.5–1.5; kriteria Outfit ZSTD: -2.0 hingga +2.0. Butir yang melanggar MNSQ dan ZSTD secara bersamaan diklasifikasikan sebagai tidak valid.

Berdasarkan Tabel 2, seluruh 20 butir memiliki nilai Pt. Corr positif, berkisar antara 0.43 hingga 0.72. Dari 20 butir, sebanyak 19 butir memenuhi kriteria Outfit MNSQ (0.5–1.5) dan Outfit ZSTD (± 2.0). Butir A10 melanggar kedua kriteria secara bersamaan (Outfit MNSQ = 2.45; Outfit ZSTD = 2.15), sehingga diklasifikasikan tidak valid. Dengan demikian, 19 dari 20 butir dinyatakan valid berdasarkan analisis konstruk model Rasch.

Analisis Unidimensionalitas

Asumsi unidimensionalitas diuji menggunakan Principal Component Analysis (PCA) of standardized residuals, sebagaimana disajikan dalam Winsteps Table 23. Prosedur ini bertujuan untuk mengevaluasi apakah, setelah varians yang dijelaskan oleh konstruk laten utama diperhitungkan, residu yang tersisa masih menunjukkan pola sistematis yang dapat mengindikasikan adanya dimensi tambahan (Reeve et al., 2007; Smith, 2002). Dalam analisis Rasch, salah satu kriteria yang umum digunakan untuk mendukung unidimensionalitas adalah bahwa proporsi raw variance explained by measures minimal mencapai 20%, dengan $\geq 40\%$ dipandang sebagai standar ideal (Misbach & Sumintono, 2014; Reeve et al., 2007). Hasil analisis unidimensionalitas disajikan pada Tabel 3.

Tabel 3. Analisis Unidimensionalitas (Winsteps Table 23.0)

Komponen Varians	Eigenvalue	Observed (%)	Expected (%)
Total raw variance in observations	35.01	100.0%	100.0%
Raw variance explained by measures	15.01	42.9%	42.5%
Raw variance explained by persons	9.14	26.1%	25.9%
Raw variance explained by items	5.88	16.8%	16.7%
Raw unexplained variance (total)	20.00	57.1%	57.5%
Unexplained variance – 1st contrast	2.81	8.0%	–
Unexplained variance – 2nd contrast	2.44	7.0%	–
Unexplained variance – 3rd contrast	2.31	6.6%	–
Unexplained variance – 4th contrast	1.94	5.5%	–
Unexplained variance – 5th contrast	1.64	4.7%	–
Essential Unidimensionality (Rasch)	–	57.4%	–

Catatan. Baris berwarna hijau = indikator utama unidimensionalitas; syarat minimum: raw variance explained by measures $\geq 20\%$; standar ideal Rasch: $\geq 40\%$; seluruh unexplained variance contrasts sebaiknya $< 10\%$.

Berdasarkan Tabel 3, proporsi raw variance explained by measures mencapai 42.9% (observasi) dan 42.5% (model). Nilai unexplained variance pada lima kontras pertama masing-masing sebesar 8.0%, 7.0%, 6.6%, 5.5%, dan 4.7%, seluruhnya di bawah ambang 10%. Nilai Essential Unidimensionality sebesar 57.4%.

Analisis Kemampuan Responden (*Person Measure*)

Estimasi kemampuan responden digunakan untuk mengelompokkan siswa ke dalam tiga tingkat kemampuan berdasarkan kriteria rata-rata $\pm$ standar deviasi, dengan mean = 0.34 logit dan SD = 1.76. Berdasarkan kriteria tersebut, kategori kemampuan dibedakan menjadi tinggi (*Measure* > 2.10 logit), sedang ($-1.42 < \text{Measure} < 2.10$), dan rendah (*Measure* < -1.42). Pendekatan klasifikasi ini sejalan dengan kerangka interpretasi kemampuan berbasis *person measure* yang umum digunakan dalam analisis Rasch pada konteks pendidikan (Suryana et al., 2022; Yudha, 2023). Hasil rinci pada tingkat individu disajikan pada Tabel 4.

Tabel 4. Kemampuan Responden dan Statistik Kesesuaian (Winsteps Table 17.1)

No	Respon den	Raw Score	N	Measure (Logit)	S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. Corr	Kategori Kemampuan
1	19P	19	20	3.48	1.05	1.08	.38	.88	.38	.13	Tinggi
2	24P	19	20	3.48	1.05	.92	.19	.38	-.11	.30	Tinggi
3	02L	18	20	2.67	.78	1.14	.43	.87	.29	.19	Tinggi
4	12L	18	20	2.67	.78	.90	-.02	.51	-.09	.38	Tinggi
5	17P	18	20	2.67	.78	1.18	.50	1.72	.89	.07	Tinggi
6	20P	17	20	2.16	.67	.80	-.43	.47	-.42	.50	Tinggi
7	04L	16	20	1.75	.61	1.23	.78	1.60	.93	.15	Sedang
8	06L	14	20	1.10	.54	.96	-.10	.77	-.33	.48	Sedang
9	10L	14	20	1.10	.54	.57	-2.03	.44	-1.24	.74	Sedang
10	28P	14	20	1.10	.54	1.22	.94	1.13	.41	.28	Sedang
11	11L	13	20	.82	.53	.81	-.82	.66	-.75	.60	Sedang
12	15L	13	20	.82	.53	.61	-1.92	.50	-1.30	.73	Sedang
13	14L	12	20	.55	.52	1.52	2.13	1.56	1.40	.09	Sedang
14	22P	12	20	.55	.52	.75	-1.21	.74	-.65	.63	Sedang
15	05L	11	20	.29	.51	.69	-1.53	.68	-.95	.68	Sedang
16	09L	10	20	.03	.51	1.21	.99	1.10	.42	.35	Sedang
17	13L	10	20	.03	.51	.91	-.38	.98	.05	.52	Sedang

No	Respon den	Raw Score	N	Measure (Logit)	S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt. Corr	Kategori Kemampuan
18	21P	10	20	.03	.51	1.29	1.30	1.21	.73	.29	Sedang
19	23P	10	20	.03	.51	1.18	.86	1.16	.60	.35	Sedang
20	01L	9	20	-.23	.51	.64	-1.81	.55	-1.61	.74	Sedang
21	03L	8	20	-.50	.52	1.52	2.07	1.61	1.64	.10	Sedang
22	29P	8	20	-.50	.52	.88	-.48	.82	-.44	.57	Sedang
23	26P	7	20	-.77	.53	1.05	.27	.99	.11	.44	Sedang
24	07L	6	20	-1.07	.55	.56	-1.93	.43	-1.51	.78	Sedang
25	16P	3	20	-2.17	.69	.68	-.69	.38	-.69	.63	Rendah
26	18P	3	20	-2.17	.69	1.42	1.01	1.04	.33	.17	Rendah
27	08L	2	20	-2.73	.81	1.37	.77	2.69	1.41	.00	Rendah
28	25P	2	20	-2.73	.81	1.48	.94	2.14	1.14	-.07	Rendah
29	27P	2	20	-2.73	.81	.65	-.54	.30	-.49	.59	Rendah
Mean		11.0	20	.34	.63	1.01	-.01	.98	.01		SD=1.76

Catatan. Baris bersisir merah = responden yang mengalami *misfit*, sedangkan nilai tebal berwarna merah = parameter yang melampaui batas kriteria *fit*.

Dari 29 responden, sebanyak 6 siswa (20.69%) dikategorikan memiliki kemampuan tinggi, 18 siswa (62.07%) berada pada kategori sedang, dan 5 siswa (17.24%) berada pada kategori rendah. Nilai rata-rata person measure yang positif, yaitu sebesar 0.34 logit, menunjukkan bahwa secara umum kemampuan siswa sedikit berada di atas rata-rata tingkat kesukaran butir. Temuan ini mengindikasikan bahwa instrumen cenderung sedikit lebih mudah daripada tingkat kesulitan ideal untuk sampel penelitian ini.

Analisis Kesesuaian Responden (*Person Fit*)

Sebanyak enam responden teridentifikasi menunjukkan pola respons yang tidak sesuai (*misfitting response patterns*), yaitu 08L (*Outfit MNSQ* = 2,69), 25P (2,14), 03L (1,61), 14L (1,56), 17P (1,72), dan 04L (1,60). Secara khusus, responden 08L dan 25P masing-masing hanya memperoleh skor mentah 2 dari 20, namun mampu menjawab benar beberapa butir yang tergolong sulit. Pola respons semacam ini umumnya berkaitan dengan perilaku menebak (*guessing behavior*) atau kurangnya perhatian saat mengerjakan tes (Bond & Fox, 2015). Responden yang mengalami misfit disajikan pada Tabel 5.

Tabel 5. Responden Misfit Berdasarkan Analisis Kesesuaian Responden (*Winsteps Table 17.1*)

No	Responden	Raw Score	Measure (Logit)	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Keterangan
1	08L	2	-2.73	1.37	.77	2.69	1.41	Misfit
2	25P	2	-2.73	1.48	.94	2.14	1.14	Misfit
3	17P	18	2.67	1.18	.50	1.72	.89	Misfit
4	04L	16	1.75	1.23	.78	1.60	.93	Misfit
5	03L	8	-0.50	1.52	2.07	1.61	1.64	Misfit
6	14L	12	0.55	1.52	2.13	1.56	1.40	Misfit

Catatan. Baris berwarna merah = responden misfit; nilai tebal = parameter yang melampaui batas kriteria *fit*; kriteria *Outfit MNSQ*: 0,5–1,5; kriteria *Infit ZSTD/Outfit ZSTD*: -2,0 hingga +2,0.

Reliabilitas Instrumen

Indeks reliabilitas instrumen diperoleh dari Winsteps Table 3.1 dan disajikan pada Tabel 6 berikut ini.

Tabel 6. Ringkasan Statistik Model Rasch (Winsteps Table 3.1)

Jumlah responden	29
Jumlah butir	20
Rata-rata skor mentah responden	11,0 / 20
Skor maksimum	19
Skor minimum	2
Rata-rata logit kemampuan responden	0,34 logit
SD logit responden	1,76
Person Reliability (Real RMSE)	0,84 (Baik)
Person Separation Index	2,32
Item Reliability (Real RMSE)	0,81 (Baik)
Item Separation Index	2,10
Cronbach's Alpha / KR-20	0,89 (Sangat Baik)
Standardized Reliability (50 items)	0,94

Catatan. Baris tebal = indeks reliabilitas utama; KR-20 > 0,80 = sangat baik (Sumintono & Widhiarso, 2015).

Berdasarkan Tabel 6, nilai KR-20 sebesar 0.89 termasuk kategori sangat baik, reliabilitas responden sebesar 0.84, dan reliabilitas butir sebesar 0.81. Person separation index sebesar 2.32 menunjukkan instrumen mampu membedakan minimal tiga tingkat kemampuan siswa, sedangkan item separation index sebesar 2.10 mengindikasikan representasi minimal dua tingkat kesukaran butir yang berbeda.

PEMBAHASAN

Tingkat Kesukaran Butir

Distribusi tingkat kesukaran butir yang membentang dari $-2,85$ hingga $+1,82$ logit menunjukkan bahwa instrumen memiliki cakupan yang memadai untuk merepresentasikan variasi kemampuan siswa. Dari 20 butir, tiga tergolong sukar (A3, A17, A20), 13 tergolong sedang, dan empat tergolong mudah. Temuan ini penting dalam konteks asesmen berbasis *deep learning* karena instrumen yang baik seharusnya tidak hanya mengukur kemampuan dasar tetapi juga mampu menangkap variasi tuntutan kognitif yang beragam (Winje & Londal, 2023). Butir A19 dengan nilai $-2,85$ logit dan tingkat keberhasilan $\sim 90\%$ mengindikasikan adanya potensi *floor effect*, sehingga layak dipertimbangkan untuk direvisi atau diganti dengan butir yang lebih menantang agar akurasi pengukuran pada kemampuan rendah dapat dipertahankan (Linacre, 2023).

Sementara itu, tiga butir sukar (A3, A17, A20) menunjukkan bahwa instrumen mencakup dimensi penalaran ilmiah tingkat tinggi yang relevan dengan tuntutan *deep learning*, yaitu penerapan dan analisis konsep IPA dalam konteks baru (Qi et al., 2023). Distribusi tingkat kesukaran yang terdokumentasi secara psikometrik ini memberikan panduan konkret bagi guru dan pengembang kurikulum dalam menyusun bank soal asesmen IPA berbasis HOTS di sekolah dasar, di mana butir dengan nilai logit tinggi dapat dijadikan prototipe pengembangan soal yang mengukur kemampuan analisis dan evaluasi dalam taksonomi Bloom yang direvisi sebagaimana menjadi inti pendekatan *deep learning* (Asikainen & Gijbels, 2017; Fullan et al., 2018).

Validitas Konstruk

Seluruh 20 butir menunjukkan nilai Pt. Corr positif (0,43–0,72), yang mengonfirmasi bahwa setiap butir bergerak searah dengan konstruk laten yang diukur, yaitu kompetensi IPA. Temuan ini konsisten dengan penelitian validasi berbasis Rasch pada asesmen sains dan STEM yang menekankan pentingnya polaritas butir positif sebagai indikator awal koherensi konstruk (Naira et al., 2024; Oliva & Blanco-Lopez, 2023). Butir A10 yang mengalami *misfit* (*Outfit* MNSQ = 2,45; ZSTD = 2,15) menunjukkan adanya pola respons yang tidak terduga, di mana siswa berkemampuan rendah menjawab benar atau siswa berkemampuan tinggi menjawab salah. Pola semacam ini umumnya disebabkan oleh ambiguitas redaksi soal, *distraktor* yang tidak berfungsi optimal, atau perilaku menebak acak (Bond & Fox, 2015; Tesio et al., 2023). Meskipun nilai Pt. Corr = 0,43 masih berada dalam rentang yang dapat diterima, nilai tersebut tidak cukup untuk mengimbangi tingkat ketidaksesuaian pola respons yang terdeteksi sehingga butir A10 perlu ditinjau ulang secara substantif pada aspek redaksi *stem*, kejelasan opsi jawaban, dan fungsi pengecoh sebelum digunakan kembali. Berbeda dengan pendekatan CTT yang hanya melaporkan daya beda dan tingkat kesukaran secara sampel-dependen, pendekatan Rasch mampu mengidentifikasi secara spesifik sumber ketidaksesuaian tersebut, sehingga memberikan bukti validitas konstruk yang lebih kuat dan dapat dipercaya (Lau & Liem, 2024).

Unidimensionalitas

Proporsi *raw variance explained by measures* sebesar 42,9% telah melampaui standar ideal analisis Rasch sebesar 40% (Misbach & Sumintono, 2014), sehingga menunjukkan bahwa sebagian besar variasi respons siswa dapat dijelaskan oleh satu konstruk utama. Nilai *Essential Unidimensionality* sebesar 57,4% semakin memperkuat kesimpulan ini. Selain itu, seluruh nilai *unexplained variance* dari lima kontras PCA berada di bawah 10%, dengan pola penurunan yang konsisten antarkontras (8,0%; 7,0%; 6,6%; 5,5%; dan 4,7%), menunjukkan bahwa sisa varians lebih cenderung berupa *noise* acak daripada struktur laten tambahan yang bermakna (Smith, 2002). Temuan ini sejalan dengan penelitian Darmana et al. (2024) dan Yudha (2023) yang juga mengonfirmasi unidimensionalitas instrumen berbasis Rasch dalam konteks pendidikan sains di Indonesia, meskipun keduanya dilakukan pada jenjang SMA dan SMP. Dengan

demikian, skor total instrumen dapat diinterpretasikan secara sah sebagai representasi dari satu konstruk kemampuan IPA yang mendasari sebagaimana disyaratkan untuk validitas pengukuran yang berkualitas tinggi (Wolfe & Smith, 2007), dan butir-butir dalam instrumen ini dapat saling dipertukarkan dalam penyusunan paket soal paralel tanpa mengubah konstruk yang diukur (Hambleton et al., 2016).

Kemampuan Responden dan Kesesuaian Responden

Distribusi kemampuan siswa menunjukkan 20,69% berkemampuan tinggi, 62,07% sedang, dan 17,24% rendah. Nilai rata-rata *person measure* yang positif (0,34 logit) mengindikasikan bahwa instrumen cenderung sedikit lebih mudah dibandingkan kemampuan rata-rata sampel, sehingga penambahan beberapa butir dengan tingkat kesukaran lebih tinggi disarankan agar ketepatan pengukuran pada siswa berkemampuan tinggi dapat ditingkatkan (Qi et al., 2023). Hal ini penting dalam konteks *deep learning* yang menuntut variasi tuntutan kognitif lebih luas (Winje & Lوندal, 2023). Dari sisi kesesuaian responden, enam siswa (20,69%) menunjukkan pola respons tidak konsisten (*misfit*). Responden 08L dan 25P, masing-masing dengan skor mentah hanya 2 dari 20, mampu menjawab benar beberapa butir sukar, yang mengindikasikan perilaku menebak (*guessing behavior*). Proporsi 20,69% masih dalam batas yang dapat diterima pada penelitian dengan sampel kecil dan tidak mengancam validitas kalibrasi butir secara keseluruhan (Linacre, 2023). Keunggulan analisis Rasch dibandingkan CTT di sini terletak pada kemampuannya mengidentifikasi secara spesifik siswa mana yang menunjukkan pola belajar tidak konsisten, sehingga guru dapat memberikan intervensi yang tepat sasaran untuk mengembangkan kebiasaan berpikir yang lebih reflektif sebagaimana dituntut oleh pendekatan *deep learning* (Quinn et al., 2019).

Reliabilitas Instrumen

Nilai KR-20 sebesar 0,89 termasuk dalam kategori sangat baik ($>0,80$) dan sebanding dengan hasil penelitian validasi berbasis Rasch yang sejenis (Hamzah et al., 2022; Scoulas et al., 2021). Koefisien *person reliability* sebesar 0,84 menunjukkan bahwa urutan kemampuan siswa cenderung akan tetap konsisten apabila instrumen diberikan dalam bentuk tes ekuivalen, sedangkan *item reliability* sebesar 0,81 mencerminkan stabilitas estimasi tingkat kesukaran butir yang dapat digeneralisasikan ke sampel lain dari populasi yang sama (Boone, 2016). Nilai *person separation index* sebesar 2,32 mengindikasikan bahwa instrumen mampu membedakan setidaknya tiga tingkat kemampuan siswa, yang merupakan karakteristik penting untuk instrumen dalam lingkungan *deep learning* di mana perbedaan kemampuan antarpeserta didik diperkirakan signifikan. Nilai *item separation index* sebesar 2,10 menunjukkan bahwa butir-butir soal merepresentasikan setidaknya dua tingkat kesukaran yang berbeda sehingga mendukung klasifikasi kemampuan siswa secara lebih efektif. Profil reliabilitas ini konsisten dengan standar kualitas instrumen berbasis Rasch dalam pendidikan sains dan menegaskan bahwa instrumen telah memenuhi persyaratan psikometrik minimum

untuk digunakan dalam konteks pengukuran kompetensi IPA di sekolah dasar (Boone, 2016; Hambleton et al., 2016).

Keterbatasan Penelitian

Penelitian ini memiliki keterbatasan utama pada ukuran sampel yang hanya melibatkan 29 siswa dari satu sekolah dasar, sehingga generalisasi hasil kalibrasi butir ke populasi yang lebih luas perlu dilakukan secara hati-hati. Nilai *standard error* (S.E.) butir yang berkisar antara 0,46 hingga 0,69 logit masih berada pada rentang yang dapat diterima untuk tujuan kalibrasi awal instrumen (Linacre, 2023). Sampel yang berasal dari satu kelas berpotensi membatasi keberagaman karakteristik responden sehingga stabilitas hierarki kesukaran butir belum dapat dipastikan akan bertahan konsisten apabila instrumen diujikan pada populasi yang lebih luas dan heterogen. Oleh karena itu, temuan ini sebaiknya dimaknai sebagai bukti awal (*preliminary evidence*) mengenai kualitas psikometrik instrumen yang masih memerlukan replikasi pada sampel yang lebih besar dan berasal dari beberapa sekolah dasar dengan latar belakang sosial-ekonomi yang beragam.

SIMPULAN

Instrumen asesmen IPA yang dikembangkan dalam kerangka *deep learning* terbukti memiliki kualitas psikometrik yang memadai, ditunjukkan oleh reliabilitas yang sangat baik, validitas konstruk yang terkonfirmasi, dan terpenuhinya asumsi unidimensionalitas, sehingga layak digunakan untuk mengukur kompetensi IPA siswa kelas V sekolah dasar. Temuan ini masih perlu direplikasi pada sampel yang lebih besar dan beragam sebelum dapat digeneralisasikan secara luas, dan penelitian ini menawarkan kerangka metodologis berbasis model Rasch yang lebih ketat dan komprehensif dibandingkan pendekatan *Classical Test Theory* untuk menjamin kualitas instrumen asesmen IPA berbasis HOTS di Indonesia.

Berdasarkan temuan penelitian ini, beberapa rekomendasi dapat dikemukakan. Pertama, butir A10 perlu direvisi untuk mengurangi ambiguitas redaksi dan meminimalkan kemungkinan respons menebak. Kedua, perlu ditambahkan beberapa butir dengan tingkat kesukaran tinggi agar akurasi pengukuran terhadap siswa berkemampuan tinggi dapat ditingkatkan. Ketiga, penelitian selanjutnya sebaiknya melibatkan jumlah sampel yang lebih besar dan lebih beragam dari beberapa sekolah dasar guna memperkuat stabilitas dan generalisasi kalibrasi butir. Keempat, pelaksanaan asesmen berikutnya disarankan untuk memasukkan analisis *Differential Item Functioning* (DIF) guna mendeteksi potensi bias butir berdasarkan variabel seperti jenis kelamin, tipe sekolah, atau perbedaan wilayah.

DAFTAR PUSTAKA

Asikainen, H., & Gijbels, D. (2017). Do students develop towards more deep approaches to learning during studies? A systematic review on the development of students' deep and surface approaches to learning in higher education. *Educational*

- Psychology Review*, 29(2), 205–234. <https://doi.org/10.1007/s10648-017-9406-6>
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences (3rd ed.)*. Routledge.
- Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE—Life Sciences Education*, 15(4), rm4. <https://doi.org/10.1187/cbe.16-04-0148>
- Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, 119(2), 166.e7–166.e16. <https://doi.org/10.1016/j.amjmed.2005.10.036>
- Darmana, A., Pulungan, A. N., Nasution, H. A., Nst, M. A., & Faradilla, P. (2024). Test instrument validation with Rasch model. In Proceedings of the 5th International Conference on Innovation in Education, Science, and Culture (ICIESC 2023). European Alliance for Innovation. <https://doi.org/10.4108/eai.24-10-2023.2342115>
- Fanani, A., & Kusmaharti, D. (2018). Pengembangan pembelajaran berbasis HOTS (higher order thinking skill) di sekolah dasar kelas V. *Jurnal Pendidikan Dasar*, 9(1), 1–11. <https://doi.org/10.21009/JPD.081>
- Fullan, M., Quinn, J., & McEachen, J. (2018). *Deep learning: Engage the world change the world*. Corwin.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (2016). *Fundamentals of item response theory*. SAGE.
- Hamzah, F. M., Rashid, M. N. A., Rahman, M. N. A., & Rasul, M. S. (2022). Evaluating the validity and reliability of authentic learning instruments using Rasch model. *International Journal of Global Optimization and Its Application*, 1(3), 182–189. <https://doi.org/10.56225/ijgoia.v1i3.69>
- Isrokatun, I., Sunaengsih, C., Maulana, M., Syahid, A. A., Karlina, D. A., & Rohaeti, P. (2024). HOTS-based learning versus HOTS-based test instruments: A study analysis of elementary school teachers' and students' readiness. *Mimbar Sekolah Dasar*, 11(3), 595–612. <https://doi.org/10.53400/mimbar-sd.v11i3.75404>
- Lau, C. K. H., & Liem, G. A. D. (2024). *Assessment approaches and deep learning*. *Educational Psychology Review*, 36(1), Article 10. <https://doi.org/10.1007/s10648-024-09856-3>
- Linacre, J. M. (2023). *Winsteps Rasch measurement computer program user's guide*. Winsteps.com. <https://www.winsteps.com>
- Misbach, I. H., & Sumintono, B. (2014). Pengembangan dan validasi instrumen "Persepsi siswa terhadap karakter moral guru" di Indonesia dengan model Rasch. Proceedings of the National Psychometrics Seminar.
- Naira, H., Abdul Aziz, S. N., & Othman, M. R. (2024). Rasch analysis of chemistry assessment. *Malaysian Journal of Innovative Education and Assessment Science (MYJIEAS)*, 4(1), 1–14.
- Oliva, J. M., & Blanco-Lopez, A. (2023). Rasch analysis and validity of the construct

- understanding of the nature of models in Spanish-speaking students. *European Journal of Science and Mathematics Education*, 11(2), 344–359. <https://doi.org/10.30935/scimath/12712>
- Postareff, L., Mattsson, M., Lindblom-Ylänne, S., & Hailikari, T. (2019). The complex relationship between emotions, approaches to learning, study success and study progression during the transition to university. *Higher Education*, 77(3), 441–457. <https://doi.org/10.1007/s10734-018-0290-x>
- Purwaningsih, N., Utaminingsih, S., & Surachmi, S. (2022). Development of assessment instruments based on higher order thinking skills (HOTS) in thematic learning of grade IV elementary school students. *Jurnal Prakarsa Paedagogia*, 5(1), 165–175. <https://doi.org/10.24176/jpp.v5i1.8674>
- Qi, H., Kim, M., Li, Y., Sandoval, C. L., Schurgers, C., Lubarda, M. V., Wagstaff, J. M., Nader, G., & Phan, A. (2023). From rote learning to deep learning: Filling the gap by enhancing engineering students' reasoning skills through explanatory learning activities. *Journal of Engineering Education*, 112(4), 1027–1048. <https://doi.org/10.1002/jee.20535>
- Quinn, J., McEachen, J., Fullan, M., Gardner, M., & Drummy, M. (2019). *Dive into deep learning: Tools for engagement*. Corwin.
- Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Yu, L., Cella, D., & Waldron, T. K. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, 45(5), S22–S31. <https://doi.org/10.1097/01.mlr.0000250483.85507.04>
- Scoulas, J. M., Aksu Dünya, B., & De Groote, S. L. (2021). Validating students' library experience survey using Rasch model. *Library and Information Science Research*, 43(1), Article 101071. <https://doi.org/10.1016/j.lisr.2021.101071>
- Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, 3(2), 205–231.
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan*. Trim Komunika.
- Suryana, D., Putri, M. A., Supriatna, M., & Yudha, E. S. (2022). Analisis Rasch model: Validitas dan reliabilitas instrumen korban bullying. *Hisbah: Jurnal Bimbingan Konseling dan Dakwah Islam*, 19(2), 113–125.
- Tesio, L., Simone, A., & Martinuzzi, G. (2023). Interpreting results from Rasch analysis 2: Advanced model applications and the data-model fit assessment. *Disability and Rehabilitation*, 45(3), 604–617. <https://doi.org/10.1080/09638288.2023.2169772>
- Wibisono, S. (2014). Aplikasi model Rasch untuk validasi instrumen pengukuran fundamentalisme agama bagi responden Muslim. *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia*, 3(3), 729–750.
- Winje, Ø., & Løndal, K. (2023). The why, what and how of deep learning: Critical

- analysis and additional concerns. *Nordic Journal of Studies in Educational Policy*, 9(2), 74–87. <https://doi.org/10.1080/20004508.2023.2194502>
- Wolfe, E. W., & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II—Validation activities. *Journal of Applied Measurement*, 8(2), 204–234.
- Yudha, R. P. (2023). Higher order thinking skills (HOTS) test instrument: Validity and reliability analysis with the Rasch model. *Eduma: Mathematics Education Learning and Teaching*, 12(1), 21–38. <https://doi.org/10.24235/eduma.v12i1.9468>