Pendas : Jurnal Ilmiah Pendidikan Dasar,
ISSN Cetak : 2477-2143 ISSN Online : 2548-6950
Volume 09 Nomor 04, Desember 2024

PENGEMBANGAN INSTRUMEN ASESMEN TES *HIGHER ORDER THINKING SKILLS* PADA PEMBELAJARAN IPAS KELAS V SD

Dewi Yuliyani[1], Supriyadi[2] , Yeri Sutopo[3]
[1,2,3] Pascasarjana Universitas Negeri Semarang
[1]dewiyuliyani14@students.unnes.ac.id ,[2] supriyadi@mail.unnes.ac.id,
[3]yerisutopo@mail.unnes.ac.id

ABSTRACT

This research aims to develop an assessment instrument to measure the high-level thinking abilities (HOTS) of fifth grade elementary school students in science learning. The instrument developed consists of 10 descriptive questions that have been validated. Data analysis using the Aiken formula and Rasch model shows that the instrument has good quality in measuring HOTS. In both small and large scale tests, the instrument was proven to be unidimensional and reliable. However, further analysis revealed the existence of several items that indicated gender bias. However, overall the instrument can be used to measure student HOTS. The results of this research contribute to the development of better assessment instruments to measure elementary school students' higher-order thinking abilities.

Keywords*: test assessment, HOTS, IPAS, rasch*

ABSTRAK

Penelitian ini bertujuan mengembangkan instrumen asesmen untuk mengukur kemampuan berpikir tingkat tinggi (HOTS) siswa kelas V SD dalam pembelajaran IPA. Instrumen yang dikembangkan terdiri dari 10 soal uraian yang telah divalidasi. Analisis data menggunakan rumus Aiken dan model Rasch menunjukkan bahwa instrumen memiliki kualitas yang baik dalam mengukur HOTS. Baik pada uji skala kecil maupun besar, instrumen terbukti unidimensional dan reliabel. Namun, analisis lebih lanjut mengungkapkan adanya beberapa item yang menunjukkan bias gender. Meskipun demikian, secara keseluruhan instrumen dapat digunakan untuk mengukur HOTS siswa. Hasil penelitian ini memberikan kontribusi dalam pengembangan instrumen asesmen yang lebih baik untuk mengukur kemampuan berpikir tingkat tinggi siswa SD.

Kata Kunci: asesmen tes, HOTS, IPAS, rasch

A. Pendahuluan

Pendidikan moderen menekankan pentingnya pengembangan keterampilan Higher Order Thinking Skills (HOTS) pada peserta didik. HOTS atau Keterampilan Berpikir Tingkat Tinggi merupakan kemampuan untuk menganalisis, mengevaluasi, dan merumuskan solusi baru dalam menghadapi permasalahan kompleks. Keterampilan ini sangatlah esensial untuk mempersiapkan generasi muda dalam menghadapi berbagai rintangan dan perubahan di masa depan.

Berbeda dengan pembelajaran tradisional yang menekankan hafalan, HOTS mendorong peserta didik untuk secara aktif menginterpretasikan, menganalisis, dan bahkan memodifikasi informasi yang mereka terima. Pendekatan ini menjadikan pembelajaran lebih menarik, interaktif, dan tidak monoton. Azam & Rokhimawan, (2020) mengatakan melatih kemampuan berpikir tingkat tinggi (HOTS) pada peserta didik melalui sistem pendidikan menjadi kunci untuk mempersiapkan generasi penerus dalam menghadapi dan beradaptasi dengan perubahan secara kreatif dan solutif.

Kurikulum Merdeka menekankan pengembangan keterampilan berpikir kritis dan kreatif (HOTS) pada peserta didik, yaitu kemampuan menalar, mengolah, dan menyajikan informasi dengan cara yang inovatif, kritis, mandiri, kooperatif, dan komunikatif. Salah satu tantangan utama dalam menerapkan kurikulum ini adalah pada aspek pembelajaran, khususnya dalam hal perubahan sistem evaluasi atau penilaian (Supriyadi et al., 2022).

Di tengah urgensi pengembangan HOTS, dunia pendidikan masih terkendala berbagai hambatan. Salah satu yang paling krusial adalah minimnya pemahaman guru tentang HOTS itu sendiri. Hal ini berakibat pada kesulitan guru dalam merancang asesmen HOTS yang efektif dan sesuai dengan tingkat kemampuan peserta didik. Kesulitan guru dalam mengembangkan instrumen asesmen HOTS ini bersumber dari beberapa faktor. Pertama, kurangnya pemahaman terhadap kata kerja operasional yang dikategorikan sebagai HOTS. Kedua, guru belum mampu menyesuaikan kompetensi dasar dan indikator dalam

penyusunan Rencana Pelaksanaan Pembelajaran (RPP). Akibatnya, peserta didik tidak terbiasa dengan soal-soal yang menuntut pemikiran tingkat tinggi (Maryono et al., 2022). Penerapan pembelajaran aktif dan HOTS yang melibatkan peserta didik secara langsung pun dapat menjadi solusi untuk menyelesaikan penilaian HOTS. Pendekatan ini terbukti mampu mendorong peserta didik untuk berpikir kritis dan kreatif, sehingga mereka dapat menjawab soal-soal HOTS dengan lebih baik (Suwarma & Apriyani, 2022).

Penerapan pembelajaran aktif dan HOTS yang melibatkan peserta didik secara langsung pun dapat menjadi solusi untuk menyelesaikan penilaian HOTS. Pendekatan ini terbukti mampu mendorong peserta didik untuk berpikir kritis dan kreatif, sehingga mereka dapat menjawab soal-soal HOTS dengan lebih baik (Purnasari et al., 2021). Instrumen asesmen HOTS yang dirancang dengan cermat dan teliti menjadi kunci untuk mengukur kemampuan berpikir tingkat tinggi peserta didik secara presisi. Akurasi instrumen asesmen ini pun secara signifikan memengaruhi hasil belajar peserta didik (Aryadi & Margunayasa, 2022). Instrumen asesmen HOTS yang dirancang dengan cermat dan teliti menjadi kunci untuk mengukur kemampuan berpikir tingkat tinggi peserta didik secara presisi. Akurasi instrumen asesmen ini pun secara signifikan memengaruhi hasil belajar peserta didik (Frisela Ratna Yuparing, Bambang Budi Wiyono, 2023).

Seiring dengan perkembangan kurikulum yang menuntut kemampuan berpikir tingkat tinggi (HOTS) pada peserta didik, maka instrumen asesmen berbasis HOTS menjadi kebutuhan penting. Hal ini mendorong guru untuk mengembangkan kemampuan dalam menyusun soal-soal HOTS, baik dalam asesmen formatif maupun sumatif. Soal-soal HOTS, yang dirancang untuk mengukur keterampilan berpikir kritis, kreatif, evaluatif, dan metakognitif (C4, C5, dan C6 dalam Taksonomi Bloom revisi), dapat melatih peserta didik dalam mengasah kemampuan dan keterampilannya sesuai dengan tuntutan kompetensi abad 21. Kurangnya soal HOTS dalam buku peserta didik IPAS kelas V dan tuntutan kurikulum merdeka yang menekankan kemampuan berpikir tingkat tinggi (HOTS) melandasi

penelitian ini untuk mengembangkan instrumen asesmen tes HOTS analisis model Rasch. Pengembangan ini diharapkan dapat memperkaya bank soal HOTS untuk penilaian dan latihan HOTS peserta didik, serta membantu guru dalam melatih kemampuan berpikir tingkat tinggi mereka.

B. Metode Penelitian

Penelitian ini menggunakan metode penelitian dan pengembangan (RnD) untuk menghasilkan instrumen asesmen keterampilan berpikir tingkat tinggi (HOTS) yang praktis untuk pembelajaran IPAS. Sugiyono (2018) mengatakan Metode penelitian dan pengembangan (RnD) merupakan pendekatan sistematis yang bertujuan menciptakan produk baru berkualitas tinggi dan inovatif. Ini dimulai dengan melakukan penelitian untuk mengumpulkan pengetahuan dasar tentang suatu subjek.

Pengetahuan ini kemudian diubah menjadi produk nyata, yang selanjutnya dievaluasi efektivitasnya untuk memastikan kesesuaiannya untuk digunakan. Pengembangan instrumen asesmen tes HOTS pada pembelajaran IPAS di kelas V Sekolah Dasar menggunakan model pengembangan Djemari Mardapi. Tahapan model pengembangan (I.P. Harysmantara et al., 2022) dibatasi dan disesuaikan dengan kebutuhan peneliti. Oleh karena itu, tidak semua langkah dari model pengembangan tersebut digunakan dalam penelitian ini. Dalam penelitian ini, hanya beberapa tahap yang digunakan, seperti melakukan analisis kebutuhan, menyusun spesifikasi tes, menulis soal tes, menelaah soal tes, melakukan uji coba tes, menganalisis butir soal tes, dan memperbaiki tes. Langkah merakit tes tidak digunakan karena langkah tersebut dilakukan saat memperbaiki tes. Sedangkan tahap melaksanakan tes tidak digunakan karena langkah tersebut sama dengan tahap uji coba. Dalam penelitian ini, uji coba hanya dilakukan sekali pada langkah kelima. Dengan asumsi bahwa hasil telaah yang dilakukan oleh para ahli mampu menjamin kualitas instrumen yang dibuat.

Subjek yang digunakan dalam penelitian ini menggunakan uji kelayakan dan uji kepraktisan. Uji kepraktisan dan uji kelayakan (validitas, reliabilitas, dan tingkat kesukaran) dilakukan pada skala

kecil dan skala besar. Jumlah peserta didik di kelas V SDN Rawu dengan 30 peserta didik dan jumlah di kelas V SDN Serang 03 dengan 100 peserta didik. Teknik observasi digunakan untuk menentukan permasalahan di lapangan melalui proses pengamatan secara langsung. Teknik observasi digunakan untuk menentukan permasalahan di lapangan melalui proses pengamatan secara langsung. Penelitian ini menggunakan teknik wawancara untuk mengumpulkan data secara langsung dari narasumber yang memiliki informasi penting terkait penelitian. Proses wawancara dilakukan dengan cara peneliti mengajukan pertanyaan kepada narasumber dan mencatat jawabannya. Pedoman wawancara telah dibuat untuk memastikan bahwa peneliti mendapatkan informasi yang relevan dengan penelitian (Inovayani Saragih, Santri Angelia Damanik, Ayi Darmana, 2023) .Teknik angket pada penelitian ini yang digunakan untuk mengetahui informasi secara langsung dari peserta didik mengenai keterbacaan instrumen asesmen tes HOTS yang diberikan. Selain itu teknik angket juga digunakan untuk mengetahui kualitas produk yang telah dikembangkan yaitu tingkat validitas yang diperoleh dari 4 validator ahli dan respon mengenai produk oleh peserta didik.

Penelitian ini menguji keabsahan soal tes dengan rumus Aiken di Microsoft Excel. Hasil perhitungan dibandingkan dengan tabel Aiken. Soal tes yang digunakan guru di kelas 5 IPAS dianalisis dengan studi dokumen. Soal tes dianggap sah jika nilainya lebih besar atau sama dengan tabel Aiken. Sebaliknya, soal tes dianggap tidak sah jika nilainya lebih kecil dari tabel Aiken. Peneliti melakukan pengujian validitas konstruk instrumen tes dengan menggunakan model Rasch. Validitas konstruk adalah validitas yang menunjukkan apakah instrumen tes mengukur apa yang seharusnya diukur. Untuk memastikan alat ukur mengukur apa yang seharusnya, peneliti menggunakan model Rasch dan melihat nilai *Item Polarity*. Nilai Point Measure Correlation (Pt.mea-Corr) yang positif menunjukkan kesesuaian antara item dan konsep yang diukur. Sedangkan nilai Mean Square Outfit (MSO) yang lebih kecil dari 1.5 menunjukkan pengukuran yang akurat. Dalam penelitian ini, semua item memiliki nilai Pt.mea-

Corr positif dan MSO di bawah 1.5. Hal ini menunjukkan bahwa semua item pada alat ukur ini mengukur konsep yang tepat dan digunakan untuk mengukur kemampuan peserta didik dengan baik.

Analisis model Rasch menggunakan Winsteps memberikan informasi yang sangat berharga dalam pengembangan instrumen penilaian. Melalui Winsteps, kita dapat mengukur reliabilitas instrumen (Person Reliability dan Item Reliability), tingkat kesulitan setiap soal, dan kesesuaian soal dengan model (item fit). Soal yang memiliki "item fit" yang baik menunjukkan bahwa soal tersebut mengukur kemampuan peserta didik secara tepat. Sebaliknya, soal yang tidak "fit" perlu diperbaiki atau dihapus karena kemungkinan besar soal tersebut tidak dipahami dengan baik oleh peserta didik. Selain itu, Winsteps juga memungkinkan kita untuk mengidentifikasi adanya bias item (DIF) berdasarkan variabel demografi seperti jenis kelamin dan domisili. Dengan demikian, analisis Rasch membantu memastikan bahwa instrumen penilaian yang kita gunakan memiliki kualitas yang baik dan dapat memberikan hasil pengukuran yang akurat. Proses validasi instrumen menggunakan data ordinal yang diperoleh dari tanggapan para ahli terhadap instrumen yang dikembangkan. Tanggapan ini digunakan untuk menyempurnakan instrumen agar menjadi lebih valid dan reliabel. Skor dari tanggapan para ahli dianalisis menggunakan Microsoft Office Excel dengan formula Aiken. Formula Aiken digunakan untuk menghitung dan menentukan validity coefficient (V) dari data berskala rating secara statistik Formula Aiken dirumuskan sebagai berikut.

$$V = \frac{S}{[n(c-1)]}$$

keterangan:
V = jumlah s dari setiap *n*
S = r-lo
r = skor yang diberikan rater (para ahli)
lo = angka penilaian validitas terendah
n = banyaknya penilai ahli (rater)
c = angka penilaian validitas tertinggi
1 = angka ketetapan sebesar 1

Rumus Aiken membantu menghitung validitas isi instrumen berdasarkan penilaian para ahli terhadap setiap butir soal. Semakin banyak ahli yang menilai (n rater), semakin tinggi nilai indeks Aiken yang dianggap valid. Untuk tujuh ahli, nilai minimal yang dianggap valid adalah 0,75, sedangkan untuk enam

ahli, nilai minimalnya adalah 0,79. Butir soal dengan nilai indeks Aiken di bawah nilai minimal tersebut dianggap tidak valid dan perlu diperbaiki atau dihapus.

Bambang Sumintono dan Wahyu Widhiarso (2015) menjelaskan bahwa dimensi merupakan komponen ukur yang unik, yang tidak memiliki hubungan satu dengan yang lainnya. Alat ukur yang mengukur satu aspek atau elemen secara unik disebut unidimensi. Uji asumsi unidimensi digunakan untuk mengetahui apakah instrumen asesmen tes HOTS hanya mengukur satu aspek atau elemen. Untuk menilai tingkat kesesuai butir soal (item fit), terdapat tiga kriteria, yaitu nilai outif means-square, outfit z-standard, dan point measure correlation.

Butir soal yang tidak memenuhi ketiga kriteria validitas, yaitu tingkat kesesuaian, daya pembeda, dan tingkat kesulitan, harus diperbaiki atau diganti. Penilaian butir soal ini penting untuk memastikan bahwa tes yang diberikan dapat mengukur kemampuan peserta didik dengan tepat. Reliabilitas dalam permodelan Rasch dapat diuji dengan menggunakan dua nilai, yaitu nilai reliability person dan item, serta nilai cronbach alpha. Nilai reliability person dan item menunjukkan tingkat konsistensi jawaban peserta didik dan butir soal. Nilai cronbach alpha menunjukkan tingkat konsistensi keseluruhan instrumen penilaian. Selain itu, reliabilitas juga dapat dilihat dari nilai separasi peserta didik dan butir soal. Nilai separasi peserta didik menunjukkan tingkat penyebaran kemampuan peserta didik.

Butir soal yang baik adalah butir soal yang netral dan tidak memihak. Peneliti melakukan analisis item DIF dengan melihat nilai probabilitas pada tabel keluaran item DIF. Nilai probabilitas yang lebih besar dari 5% atau 0,05 menunjukkan bahwa butir soal tersebut tidak bias. Model Rasch dan program Winstep membantu peneliti menganalisis tingkat kesukaran butir soal. Model Rasch menghasilkan nilai logit untuk setiap soal, di mana nilai yang lebih tinggi menunjukkan soal yang lebih sulit. Tingkat kesukaran soal berdasarkan jawaban peserta didik dapat dilihat pada nilai measure dalam tabel keluaran Winstep. Soal yang baik adalah soal yang dapat mengukur kemampuan peserta didik

dan membedakan kemampuan antar peserta didik. Kemampuan soal dalam membedakan kemampuan peserta didik dapat dilihat dari nilai standard error (SE). Nilai SE yang kurang dari 0,5-1,00 menunjukkan soal yang baik atau ideal. Peneliti juga dapat melihat banyaknya kelompok kesulitan butir soal dengan nilai separasi atau kelompok soal. Kelompok-kelompok ini dapat dilihat pada tabel keluaran summary statistics Winstep.

C.Hasil Penelitian dan Pembahasan

Pada uji coba skala kecil pengujian unidimensi bertujuan untuk mengatahui apakah instrumen asesmen tes HOTS yang dikembangkan hanya mengukur satu kemampuan yaitu kemampuan berpikir tingkat tinggi pada peserta didik atau tidak. Jika uji asumsi unidimensi terpenuhi maka analisis data menggunakan model Rasch dapat dilanjutkan. Bambang Sumintono dan Wahyu Widhiarso(2015) mengemukakan bahwa persyaratan minimal asumsi unidimensional yaitu sebesar 20% dan unidimensional Rasch sebesar 40%. Uji asumsi unidimensi pada tahap skala kecil dapat dilihat pada Tabel 1 berikut.

Tabel 1. Standar Residual Varians Uji Coba Skala Kecil

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = Item information units

		Eigenvalue	Observed		Expected
Total raw variance in observations	=	15.2978	100.0%		100.0%
Raw variance explained by measures	=	5.2978	34.6%		34.5%
Raw variance explained by persons	=	1.7630	11.5%		11.5%
Raw Variance explained by items	=	3.5348	23.1%		23.1%
Raw unexplained variance (total)	=	10.0000	65.4%	100.0 %	65.4%
Unexplned variance in 1st contrast	=	2.1623	14.1%	21.6 %	
Unexplned variance in 2nd contrast	=	1.6306	10.7%	16.3 %	
Unexplned variance in 3rd contrast	=	1.5388	10.1%	15.4 %	
Unexplned variance in 4th contrast	=	1.4012	9.2%	14.0 %	
Unexplned variance in 5th contrast	=	1.0756	7.0%	10.8 %	

Berdasarkan data pada tabel 1 terlihat bahwa pengukuran raw variance explained by measures bernilai 34,5%. Hal ini menunjukkan bahwa persyaratan nilai unidimensinya telah terpenuhi. Nilai minimal yang disyaratkan adalah 20% sedangkan jika di atas 60% adalah istimewa. Nilai lain yang perlu dilihat adalah unexplained variance. Nilai unexplained variance pada observed idealnya tidak melebihi 15%. Hasil yang diperoleh terlihat bahwa semua nilai observed unexplained variance nya di bawah 15%. Hal ini membuktikan bahwa unidimensionalitas dari instrumen asesmen tes baik. Dalam Permodelan Rasch selain nilai reliabilitas Alpha Cronbach, juga

memunculkan dua jenis reliabilitas person dan reliabilitas item. Hasil pengujian reliabilitas dapat dilihat pada Tabel 2 berikut.

Tabel 2. Reabilitas Person dan Item Uji Skala Kecil

	Person	Item
Reliabilitas	0,62	0,83
Separation	1,28	2,19
Mean	0,16	0,23
SD	0,69	0,64
Outfit MNSQ	1,00	1,00
Outfit ZSTD	0,02	-0,06
Cronbach Alpha (0,65)		

Berdasarkan pada tabel dapat dilihat bahwa nilai reliabilitas cronbach alpha adalah 0,65 yang menunjukan nilai reliabilitas yang cukup. Hal ini menunjukan bahwa interaksi antar persin dan item yang cukup. Nilai reliabilitas person adalah 0,62 yang menunjukan konsistensi jawaban dari person cukup. Nilai separation personnya adalah 1,28 yang menunjukan nilai pengelompokan person yang kecil, artinya person kurang bervariasi dalam pengelompokan. Nilai reliabilitas item bagus dengan nilai 0,83 dengan nilai separtion sebesar 2,19 yang menunjukan item-item memiliki variasi tingkat kesulitan yang cukup bagus. Pengujian item fit ini diperlukan untuk mengetahui apakah suatu item berfungsi normal dalam melakukan pengukuran (Bambang Sumintono dan Wahyu Widhiarso, 2015). Hasil pengukuran item fit dapat dilihat pada Tabel 3 berikut.

Tabel 3. Item Fit Uji Skala Kecil

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
1	58	32	.10	.27	1.31	1.47	1.43	1.72	A .35	.50	46.9	49.2	item1
4	63	32	-.25	.26	1.27	1.32	1.35	1.49	B .35	.51	40.6	49.8	item4
7	61	32	-.11	.26	1.20	1.03	1.10	.52	C .67	.51	34.4	49.1	item7
10	67	32	-.53	.26	1.16	.84	1.14	.68	D .44	.51	50.0	49.8	item10
9	55	32	.32	.27	1.07	.41	.95	-.13	E .61	.49	46.9	52.6	item9
3	72	32	-.88	.27	1.01	.12	1.06	.31	e .53	.50	40.6	50.6	item3
2	41	32	1.63	.37	.83	-.44	.94	.03	d .32	.38	71.9	75.7	item2
6	50	32	.71	.29	.82	-.75	.88	-.32	c .34	.47	56.3	57.2	item6
8	69	32	-.67	.27	.62	-2.11	.62	-1.87	b .48	.51	71.9	50.5	item8
5	64	32	-.32	.26	.57	-2.52	.55	-2.36	a .76	.51	65.6	49.7	item5
MEAN	60.0	32.0	.00	.28	.99	-.1	1.00	.0			52.5	53.4	
P.SD	8.9	.0	.70	.03	.25	1.3	.27	1.2			12.7	7.8	

Pengujian item fit menggunakan 3 kriteria yaitu outft MNSQ, ZSTD dan Point measure correlation

• Outfit MNSQ yang diterima: 0,5 < MNSQ < 1,5

• Outfit ZSTD yang diterima: -2 < ZSTD < 2

• Outfit Point Measure Correlation yang diterima: 0,4 < PT Mea Corr < 0.8

Berdasarkan ketiga kriteria ini maka seluruh item memenuhi kriteria, artinya bahwa seluruh item fit dan dapat digunakan untuk penelitian. Pendeteksian bias butir dapat menggunakan dua kriteria yaitu nilai probabilitas Mantel Chi Square dan nilai DIF Contrast. Jika nilai probabilitas < 0,05 maka item signifikan mengandung bias. Namun, signifikansi ini perlu dibandingkan lagi dengan nilai DIF contrast. Ada 3 kriteria DIF Contrast yaitu: Negligible, slight to moderate (DIF ≥ 0,43 logit), moderate to large (DIF ≥ 0,64 logit).

Hasil pengujian dapat dilihat pada Tabel 4 di bawah ini.

Tabel 4. Differential Item Function (DIF) Uji Coba Skala Kecil

Soal	Mantel hanzel probability	DIF contrast	Kategori DIF
1	0,1321	0,87	
2	0,6419	0,43	
3	0,2457	-0,66	
4	0,3657	-0,25	
5	0,3793	-0,46	
6	0,7079	0,00	
7	0,1610	0,83	
8	0,1157	0,69	
9	0,4566	-0,79	
10	0,1675	-0,48	

Berdasarkan tabel di atas dapat dilihat bahwa pada nilai Mantel Hanzel probability, semua item memiliki nilai probabilitas di atas 0,05 yang menandakan bahwa semua item dapat dikatakan tidak teridentifikasi bias. Namun, jika melihat pada nilai DIF contrast, maka ada beberapa item soal yang masuk kriteria DIF yaitu: item soal 2, 5 dan 10 berada pada kriteria slight to moderate DIF, item soal 1, 3, 7, 8 dan 9 berada pada kriteria moderate to large DIF. 2 item yaitu 4 dan 6 dianggap negligible atau tidak mengalami bias.

Kategorisasi item dilakukan untuk mengetahui tingkat kesulitan tiap item. Pengkategorisasian item menggunakan nilai mean dan standar deviasi dari nilai measure. Berdasarkan data, nilai mean item adalah 0,23 dan standar deviasinya adalah 0,64 dapat dilihat pada Tabel 5 berikut.

Tabel 5. Kategorisasi Item Uji Coba Skala Kecil

Keterangan	Kategorisasi	Item
SANGAT MUDAH	< -0,64	N3, N8
MUDAH	-0,64 ≤ X ≤ 0,64	N4, N7, N10, N5
SULIT	0,0 < X ≤ 0,64	N1, N9
SANGAT SULIT	X > 0,64	N2, N6

Berdasarkan tabel dapat dilihat bahwa item dengan kategori sangat mudah adalah item N3, N8. Item dengan kategori mudah adalah item N4, N7, N10 dan N5. Item dengan kategori sulit adalah item N1, N9. Dan item dengan kategori sangat sulit adalah item N2, N6. Pelaksanaan uji keterbacaan dilakukan setelah mengerjakan instrumen asesmen tes HOTS. Adapun hasil dari penilaian keterbacaan instrumen asesmen tes HOTS menunjukkan bahwa secara keseluruhan instrumen dapat dipahami dengan baik oleh 80% peserta didik. Namun, terdapat beberapa kendala yang perlu diperhatikan, seperti kurang jelasnya petunjuk pengerjaan, pertanyaan yang membingungkan, penggunaan istilah yang sulit dipahami, ketidaksesuaian pertanyaan dengan deskripsi, adanya pertanyaan dengan lebih dari satu jawaban benar, kualitas gambar yang kurang baik,

dan tingkat kesulitan soal yang tidak merata. Temuan ini mengindikasikan perlunya perbaikan pada beberapa butir soal untuk meningkatkan kualitas instrumen dan memastikan bahwa instrumen tersebut dapat mengukur kemampuan berpikir tingkat tinggi peserta didik secara akurat.

Pada uji skala besar Alat ukur kualitas hidup dalam penelitian ini relatif memiliki ukuran unidimensionalitas yang baik. Hasilnya dapat dilihat pada Tabel 6 berikut.

Tabel 6. Standardized Residual Variance Uji Skala Besar

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = Item information units

	Eigenvalue	Observed	Expected
Total raw variance in observations =	15.3556	100.0%	100.0%
Raw variance explained by measures =	5.3556	34.9%	34.8%
Raw variance explained by persons =	2.7710	18.0%	18.0%
Raw Variance explained by items =	2.5846	16.8%	16.8%
Raw unexplained variance (total) =	10.0000	65.1% 100.0%	65.2%
Unexplned variance in 1st contrast =	2.0029	13.0%	20.0%
Unexplned variance in 2nd contrast =	1.8853	12.3%	18.9%
Unexplned variance in 3rd contrast =	1.2664	8.2%	12.7%
Unexplned variance in 4th contrast =	1.1911	7.8%	11.9%
Unexplned variance in 5th contrast =	.9920	6.5%	9.9%

Dari tabel terlihat bahwa pengukuran raw variance explained by measures bernilai 43,8%. Hal ini menunjukkan bahwa persyaratan unidimensionalitasnya bagus. Nilai minimal yang disyaratkan adalah 20% sedangkan jika di atas 60% adalah istimewa. Nilai lain yang perlu dilihat adalah unexplained variance. Nilai unexplained variance pada observed idealnya tidak melebihi 15%. Hasil yang diperoleh terlihat bahwa semua nilai observed unexplained variancenya di bawah 15%. Hal ini berarti bahwa instrument dapat mengukur kualitas hidup secara efektif.

Pemodelan rasch selain mengeluarkan nilai reliabilitas Alpha Cronbach, juga memunculkan dua jenis reliabilitas sekaligus yaitu reliabilitas person dan reliabilitas item. Hasil pengujian reliabilitas dapat dilihat pada Tabel 7 berikut.

Tabel 7. Reabilitas Person dan Item Uji Skala Besar

	Person	Item
Reliabilitas	0,73	0,86
Separation	1,65	2,46
Mean	0,14	0,00
SD	1,20	0,51
Outfit MNSQ	0,99	0,99
Outfit ZSTD	-0,01	-0,24
Cronbach Alpha (0,76)		

Berdasarkan tabel dapat dilihat bahwa nilai reabilitas cronbach alpha adalah 0,76 yang menunjukkan nilai reliabilitas yang bagus. Hal ini menunjukkan bahwa interakasi antara person dan item yang bagus. Nilai reliabilitas person adalah 0,73 yang menunjukkan konsistensi jawaban dari person cukup bagus. Nilai separation personnya adalah 1,65 yang menujukkan nilai pengelompokan person yang kecil

artinya person kurang bervariasi dalam pengelompokan. Nilai relibilitas item bagus dengan nilai 0,86 dengan nilai separation sebesar 2,46 yang menujukkan item-item memiliki variasi tingkat kesulitan yang cukup bagus.

Pada rasch model, tiap item dapat diuji tingkat kecocokan modelnya. Item yang cocok dengan model Rasch dikatakan sebagai item yang fit dengan model. Pengujian item fit ini diperlukan untuk mengetahui apakah suatu item berfungsi normal dalam melakukan pengukuran. Hasil pengukuran item fit pada uji skala besar dapat dilihat pada Tabel 8 berikut.

Tabel 8. Misfit Order Uji Skala Besar

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
6	205	101	.05	.18	1.80	5.25	1.75	5.00	A .63	.56	26.7	60.3	n6
2	222	101	-.50	.18	1.27	2.05	1.28	2.10	B .26	.56	54.5	60.8	n2
5	205	101	.05	.18	1.21	1.64	1.22	1.72	C .53	.56	57.4	60.3	n5
7	226	101	-.63	.18	1.01	.16	1.00	.01	D .63	.56	62.4	61.6	n7
10	214	101	-.24	.18	.96	-.26	.94	-.44	E .72	.56	57.4	60.0	n10
4	217	101	-.34	.18	.89	-.92	.87	-1.07	e .70	.56	65.3	60.7	n4
8	191	101	.49	.18	.81	-1.58	.80	-1.62	d .61	.56	69.3	61.8	n8
1	169	101	1.23	.19	.70	-2.53	.76	-1.87	c .49	.54	73.3	63.8	n1
3	210	101	-.11	.18	.72	-2.44	.73	-2.39	b .43	.56	77.2	60.3	n3
9	206	101	.01	.18	.58	-4.01	.60	-3.80	a .58	.56	76.2	60.7	n9
MEAN	206.5	101.0	.00	.18	1.00	-.3	.99	-.2			62.0	61.0	
P.SD	15.7	.0	.51	.00	.34	2.6	.32	2.4			14.0	1.1	

Pengujian item fit menggunakan 3 kriteria yaitu outft MNSQ, ZSTD dan Point measure correlation

• Outfit MNSQ yang diterima: 0,5 < MNSQ < 1,5

• Outfit ZSTD yang diterima: -2 < ZSTD < 2

• Outfit Point Measure Correlation yang diterima: 0,4 < PT Mea Corr < 0.8

Berdasarkan ketiga kriteria ini maka dapat dikatakan ada satu item yang tidak memenuhi tiga kriteria ini yaitu item n6. Item n6 merupakan item yang terdeteksi tidak fit, karena nilai outfit MNSQ nya di atas 1,5. Nilai MNSQ menjadi nilai yang lebih diutamakan untuk dilihat dalam menilai suatu item yang fit dan tidak fit. Item ini dapat dibuang dari daftar item dan tidak diikut sertakan dalam pengukuran selanjutnya atau dengan pertimbangan tertentu dapat tetap diikutsertakan.

Pendeteksian bias butir dapat menggunakan dua kriteria yaitu nilai probabilitas Mantel Chi Square dan nilai DIF Contrast. Jika nilai probabilitas < 0,05 maka item signifikan mengandung bias. Namun, signifikansi ini perlu dibandingkan lagi dengan nilai DIF contrast. Ada 3 kriteria DIF Contrast yaitu: Negligible, slight to moderate (DIF ≥ 0,43 logit), moderate to large (DIF ≥ 0,64 logit). Hasil pengujian dapat dilihat pada Tabel 9 di bawah ini.

Tabel 9. Differential Item Function (DIF) Uji Skala Besar

Soal	Mantel hanzel probability	DIF contrast	Kategori DIF
1	0,0304	-0,75	Moderate to large
2	0,1692	0,35	Negligible
3	0,2978	0,25	Negligible
4	0,7449	0,00	Negligible
5	0,0779	0,46	Negligible
6	0,0297	-0,57	Slight to moderate
7	0,9428	-0,05	Negligible
8	0,0473	-0,53	Slight to moderate
9	0,0109	0,65	Moderate to large
10	0,9003	0,11	Negligible

Berdasarkan tabel dapat dilihat bahwa ada 2 item yang masuk kriteria DIF kategori slight to moderate yaitu item 6 dan 8, dan 2 item masuk kriteria DIF kategori moderate to large yaitu item 1 dan 9. 6 item masuk kriteria negligible atau tidak mengalami bias, yaitu item 2,3,4,5,7 dan 10. Kategorisasi item dilakukan untuk mengetahui tingkat kesulitan tiap item. Pengkategorisasian menggunakan nilai mean dan standar deviasi dari nilai measure. Berdasarkan data, nilai mean item adalah 0,00 dan standar deviasinya adalah 0,51 dapat dilihat pada tabel 10 berikut.

Tabel 10. Kategorisasi Item Uji Skala Besar

Keterangan	Kategorisasi	Item
SANGAT MUDAH	< -0,51	N7
MUDAH	-0,51 ≤ X ≤ 0,51	N2, N4, N10, N3
SULIT	0,0 < X ≤ 0,51	N9, N6, N5, N8
SANGAT SULIT	X > 0,51	N1

Berdasarkan tabel dapat dilihat bahwa item dengan kategori sangat mudah adalah item N7. Item dengan kategori mudah adalah item N2,N4,N10 dan N13. Item dengan kategori sulit adalah item N9,N6,N5,N8. Dan item dengan kategori sangat sulit adalah item N1. Pada pelaksanaan uji keterbacaan skala besar dilakukan setelah mengerjakan instrumen asesmen tes HOTS pada uji skala besar. Berdasarkan hasil angket yang diisi oleh 100 peserta didik pada uji skala besar, secara keseluruhan tingkat keterbacaan instrumen asesmen HOTS mencapai 84%. Meskipun demikian, masih terdapat beberapa kendala yang perlu diperhatikan. Banyak peserta didik merasa kesulitan memahami beberapa soal, terutama yang berkaitan dengan menyusun menu makanan sehat, perbedaan tinggi badan, dan upaya menjaga kesehatan. Selain itu, beberapa peserta juga menganggap beberapa pertanyaan terlalu mudah dan memiliki lebih dari satu jawaban yang benar. Hal ini mengindikasikan adanya beberapa soal yang perlu direvisi untuk meningkatkan kejelasan dan kesesuaian dengan tujuan penilaian.

D. Kesimpulan

Berdasarkan rumusan masalah, hasil analisis dan pembahasan data, maka diperoleh kesimpulan yang dapat diambil dari penelitian ini adalah bahwa validitas isi instrumen asesmen tes HOTS telah diuji oleh 4 ahli. 10 soal uraian dalam instrumen ini dinilai berdasarkan 3 aspek, yaitu: materi, konstruk, dan bahasa. Penilaian ini mengacu pada 14 kaidah yang terdapat dalam lembar penilaian validator. Data penilaian dianalisis menggunakan rumus Aiken dengan bantuan Microsoft Excel. Hasil analisis menunjukkan bahwa nilai validitas setiap butir soal pada masing-masing kaidah mencapai minimal 0,5. Hal ini menunjukkan bahwa butir-butir soal memiliki kriteria validitas sedang dan tinggi. Berdasarkan penilaian ahli, 10 soal yang dikembangkan telah memenuhi kriteria validitas.

Pengujian unidimensi uji skala kecil menunjukkan bahwa instrumen asesmen tes ini mengukur satu kemampuan yaitu kemampuan berpikir tingkat tinggi pada peserta didik. Hal ini dibuktikan dengan nilai raw variance explained by measures sebesar 34,5% (di atas 20%) dan nilai unexplained variance yang di bawah 15% untuk semua item. Reliabilitas instrumen asesmen tes ini cukup baik. Nilai reliabilitas Cronbach alpha sebesar 0,65 menunjukkan interaksi antar peserta dan item yang cukup baik.

Nilai reliabilitas person sebesar 0,62 menunjukkan konsistensi jawaban dari peserta cukup. Nilai separation personnya 1,28 menunjukkan nilai pengelompokan peserta yang kecil, artinya peserta kurang bervariasi dalam pengelompokan. Nilai reliabilitas item bagus dengan nilai 0,83 dan nilai separation sebesar 2,19 menunjukkan item-item memiliki variasi tingkat kesulitan yang cukup bagus. Semua item fit dengan model Rasch dan dapat digunakan untuk penelitian. Hal ini dibuktikan dengan terpenuhinya ketiga kriteria Outfit MNSQ, ZSTD, dan Point Measure Correlation untuk semua item.

Terdapat beberapa item yang menunjukkan bias gender. Item soal 2, 5, dan 10 berada pada kriteria slight to moderate DIF, item soal 1, 3, 7, 8, dan 9 berada pada kriteria moderate to large DIF. Item 4 dan 6 dianggap negligible atau tidak mengalami bias. Item-item dikategorikan berdasarkan tingkat

kesulitannya. Item dengan kategori sangat mudah adalah item N3 dan N8. Item dengan kategori mudah adalah item N4, N7, N10, dan N5. Item dengan kategori sulit adalah item N1 dan N9. Item dengan kategori sangat sulit adalah item N2 dan N6.

Pada uji skala besar, memiliki unidimensionalitas yang baik. Hal ini dibuktikan dengan nilai raw variance explained by measures sebesar 43,8% (di atas 20%) dan nilai unexplained variance yang di bawah 15% untuk semua item. Reliabilitas alat ukur kualitas hidup ini cukup baik. Nilai reliabilitas Cronbach alpha sebesar 0,76 menunjukkan interaksi antara person dan item yang bagus. Nilai reliabilitas person sebesar 0,73 menunjukkan konsistensi jawaban dari person cukup bagus. Nilai separation personnya 1,65 menunjukkan nilai pengelompokan person yang kecil artinya person kurang bervariasi dalam pengelompokan. Nilai reliabilitas item bagus dengan nilai 0,86 dan nilai separation sebesar 2,46 menunjukkan item-item memiliki variasi tingkat kesulitan yang cukup bagus.

Terdapat satu item yang tidak fit dengan model Rasch yaitu item n6. Hal ini dibuktikan dengan nilai outfit MNSQ nya di atas 1,5. Item ini dapat dibuang dari daftar item atau tetap diikutsertakan dengan pertimbangan tertentu. Terdapat 4 item yang menunjukkan bias gender. Item 6 dan 8 masuk kriteria DIF kategori slight to moderate, item 1 dan 9 masuk kriteria DIF kategori moderate to large. 6 item lain tidak mengalami bias. Item-item dikategorikan berdasarkan tingkat kesulitannya. Item dengan kategori sangat mudah adalah item N7. Item dengan kategori mudah adalah item N2,N4,N10 dan N13. Item dengan kategori sulit adalah item N9,N6,N5,N8. Item dengan kategori sangat sulit adalah item N1.

Instrumen asesmen tes HOTS telah diuji coba pada skala kecil dan besar, dengan hasil yang menunjukkan tingkat keterbacaan yang baik, yaitu 80% untuk uji skala kecil dan 84% untuk uji skala besar. Namun, terdapat beberapa aspek yang perlu mendapat perhatian untuk meningkatkan kualitas instrumen, khususnya terkait dengan pemahaman peserta didik.

Berdasarkan hasil uji keterbacaan, instrumen asesmen tes

HOTS perlu direvisi. Revisi ini perlu dilakukan dengan mempertimbangkan masukan dari peserta didik. Perbaikan instrumen asesmen tes HOTS dilakukan secara menyeluruh berdasarkan temuan dari uji coba skala kecil dan besar. Perubahan yang dilakukan diharapkan dapat meningkatkan kualitas instrumen dan membuatnya lebih efektif dalam mengukur kemampuan berpikir tingkat tinggi (HOTS) dalam pembelajaran IPAS.

DAFTAR PUSTAKA

Aryadi, K. S., & Margunayasa, I. G. (2022). Instrumen Penilaian High Order Thinking Skills (HOTS) pada Pembelajaran IPA. *Indonesian Journal of Instruction*, *3*(1), 34–41. https://doi.org/10.23887/iji.v3i1.44761

Azam, I. F., & Rokhimawan, M. A. (2020). Analisis Materi Ipa Kelas Iv Tema Indahnya Kebersamaan Dengan Hots. *Jurnal Ilmiah Didaktika Media Ilmiah Pendidikan Dan Pengajaran*, *21*(1), 100. https://doi.org/10.22373/jid.v21i1.5970

Bambang Sumintono dan Wahyu Widhiarso. (2015). *Aplikasi Permodelan RASCH pada assessment pendidikan*. Trim Komunikata.

Frisela Ratna Yuparing, Bambang Budi Wiyono, E. S. (2023). Instrumen Asesmen Higher Order Thinking Skills (HOTS) Pada Mata Pelajaran Matematika Kelas IV SD Negeri 4 Tanggung. *Jurnal Pendidikan Indonesia*, *4*.

I.P. Harysmantara, I.B.P. Arnyana, & I.G. Astawan. (2022). Pengembangan Instrumen Kecerdasan Naturalis Dan Kemampuan Berpikir Kritis Pada Pelajaran Ipa Kelas Iv Sekolah Dasar. *PENDASI: Jurnal Pendidikan Dasar Indonesia*, *6*(2), 35–45. https://doi.org/10.23887/jurnal_pendas.v6i2.882

Inovayani Saragih, Santri Angelia Damanik, Ayi Darmana, R. D. S. (2023). Pengembangan Instrumen Tes Berbasis HOTS Pada Materi Asam-Basa Menggunakan Model Rasch. *UNESA Journal of Chemical Education*, *Vol.12,No.*, 217–224.

Maryono, M., Sastrawati, E., & Budiono, H. (2022). Analisis Kesulitan Guru Sekolah Dasar Dalam Mengembangkan Instrumen Asesmen Higher Order Thingking Skills. *Primary: Jurnal Pendidikan Guru Sekolah Dasar*, *11*(5), 1529. https://doi.org/10.33578/jpfkip.v11i5.9182

Purnasari, P. D., Silvester, S., & Lumbantobing, W. L. (2021). Pengembangan Instrumen Asesmen Higher Order Thingking Skills (Hots) Ditinjau Dari Gaya Belajar Siswa. *Sebatik*, *25*(2), 571–580. https://doi.org/10.46984/sebatik.v

25i2.1607

Sugiyono. (2018). *Metode Penelitian Kualitatif, Kuantitatif dan R&D*. Alfabeta.

Supriyadi, S., Lia, R. M., Rusilowati, A., Isnaeni, W., Susilaningsih, E., & Suraji, S. (2022). Penyusunan Instrumen Asesmen Diagnostik untuk Persiapan Kurikulum Merdeka. *Journal of Community Empowerment*, *2*(2), 67–73. https://doi.org/10.15294/jce.v2i2.61886

Suwarma, I. R., & Apriyani, S. (2022). *Explore Teachers ' Skills in Developing Lesson Plan and Assessment That Oriented on Higher Order Thinking Skills (HOTS)*. *3*(2), 106–113. https://doi.org/10.46843/jiecr.v3i2.66