Didaktika: Jurnal Kependidikan, Vol. 13, No. 2, Mei 2024

Analisis Instrumen Penilaian Hasil Belajar Kimia untuk Mengukur *High Order Thinking Skill* (HOTS) Semester Genap Kelas X Berdasarkan Permodelan Rasch

Oddy Azis Saputra[1], Nurfajriani[2], Ajat Sudrajat[3]
[1,2,3]Universitas Negeri Medan, Indonesia

[1]nurfajriani@unimed.ac.id

Abstrak

Penelitian ini bertujan untuk mendeskripsikan hasil analisis instrumen penilaian berbasis HOTS yang dapat digunakan dalam evaluasi hasil belajar siswa di semester genap kelas X SMA. Analisis hasil instrumen ini dilakukan berdasarkan tahapan kedua dan ketiga dari model ADDIE yaitu (*design and development*). Metode yang digunakan dalam penelitian ini adalah mengumpulkan data atau informasi dari para validator ahli untuk menentukan valid atau tidak valid terhadap butir soal yang dikembangkan. Partisipan penelitian ini adalah 3 validator ahli yang berasal dari dosen dan 5 validator ahli yang berasal dari guru kimia serta 70 siswa untuk mengerjakan butir soal berbasis HOTS. Pengumpulan data dilakukan dengan menggunakan kuesioner dengan jenis data kualitatif dan menggunakan *winstep* sebagai permodelan *rasch* dengan jenis data kuantitatif. Hasil penelitian dianalisis secara deskriptif dan diperoleh data bahwa instrumen penilaian hasil belajar kimia semester genap kelas X yang dikembangkan telah dikategorikan untuk layak digunakan sebagai instrumen penilaian hasil belajar berdasarkan validasi isi dengan nilai rata-rata dari validator ahli dosen dan guru sebesar 3,74 dan 3,80. Kemudian, hasil analisis item menggunakan permodelan *rasch* dihasilkan 45 item fit dengan model, reliabilitas instrumen berada pada kategori baik dengan nilai 0,71; tingkat kesukaran item yang mendominasi terkategori sedang dengan perolehan 86,7%; daya pembeda yang mampu mendiskriminasi sebanyak 71%, dan pengecoh 77% berfungsi dengan baik. Sehingga dihasilkan 32 dari 45 soal layak dan terkategori sebagai soal yang baik.
Kata kunci: *instrumen, penilaian, HOTS, rasch*

Pendahuluan

Sebuah keberhasilan dalam pendidikan yang tujuan utamanya meningkatkan sumber daya manusia, dipengaruhi oleh berbagai faktor. Salah satu faktor yang ikut mempengaruhi keberhasilan ini adalah kemampuan guru dalam melakukan dan memanfaatkan penilaian, evaluasi proses, dan hasil belajar (Budiman & Jailani, 2014). Penilaian merupakan bagian yang tidak dapat dipisahkan dari proses belajar mengajar dan merupakan komponen penting bagi kurikulum. Dalam proses pelaksanaannya terdapat pergantian kurikulum yang menuntut perubahan orientasi pembelajaran kimia. Namun demikian, cara pembelajaran kimia tidak banyak mengalami perubahan. Secara umum, pembelajaran kimia di SMA masih didominasi oleh penyampaian informasi atau ceramah dari guru, pemberian contoh-contoh, dan latihan soal-soal. Orientasi belajar kimia pun berubah dari yang semestinya berupa penguasaan terhadap konsep-konsep ilmu kimia dan penerapannya dalam kehidupan ke latihan soal-soal ujian, baik itu untuk ujian sekolah, ujian nasional, maupun ujian masuk perguruan tinggi.

Konsep kimia dianggap kompleks karena mengharuskan untuk menguasai pengetahuan dasar sekaligus penerapan dalam kehidupan sehari-hari (Nurfajriani dkk, 2021). Banyak konsep-konsep kompleks dalam kimia yang tidak dapat hanya dijelaskan secara lisan serta memiliki

konsep yang rumit untuk dipahami oleh siswa (Yudha dkk, 2023). Dengan begitu, materi kimia memerlukan keterampilan berpikir tingkat tinggi.

Muchlis & Andromeda (2020) menyatakan bahwa kemampuan peserta didik di Indonesia khususnya di bidang kimia masih sangat rendah dibandingkan dengan negara lain. Ilmu kimia merupakan bagian dari sains. Sedangkan tes PISA (*Program for International Student Assessment*) merupakan serangkaian tes yang salah satunya menguji kemampuan literasi, matematika dan sains. Hal ini dibuktikan dengan studi internasional yaitu tes PISA yang diselenggarakan oleh OECD (*Organization for Economic Co-operation and Development*). Hasil PISA 2018 menunjukkan bahwa Indonesia berada di peringkat 69 dari 76 negara (OECD, 2019). Hasil tersebut menunjukan bahwa Indonesia masih jauh tertinggal dari beberapa Negara lainnya, dimana peserta didik di Indonesia masih kurang sekali yang dapat menjawab soal yang mengukur keterampilan berpikir tingkat tinggi dan hanya mampu menjawab soal yang tergolong dalam kategori rendah (Hanifah, 2019).

Menurut Chairani & Nurfajriani (2020) salah satu tujuan dari pendidikan nasional agar meningkatkan kemampuan, salah satu kemampuan yang ditingkatkan yakni kemampuan berpikir kritis dan kreatif. Proses pengembangan cara berpikir peserta didik dalam memecahkan suatu masalah membutuhkan pelatihan sejak dini sehingga kemampuan berpikir peserta didik akan berkembang dan memiliki kemampuan berpikir tingkat tinggi (HOTS) (Nurwahidah, 2018). Namun, masih ditemukan kesulitan yang dihadapi peserta didik dalam pembelajaran, seperti rendahnya prestasi dan capaian pada kompetensi. Solusi yang dapat ditawarkan untuk melatih kemampuan berpikir peserta menjadi lebih tinggi yaitu dengan menggunakan soal sebagai bahan latihan yang merangsang peserta didik berpikir lebih kompleks dan maju.

Salah satu taksonomi yang dikenal dalam pendidikan adalah Bloom. Fungsi Taksonomi Bloom merupakan kerangka berpikir pencapaian tujuan pembelajaran guru dalam menganalisis mata pelajaran dan membelajarkan dimensi pengetahuan serta dimensi proses kognitif yang akan dicapai oleh peserta didik. Menurut Taksonomi Bloom yang telah direvisi, proses kognitif dibedakan menjadi dua yaitu keterampilan berpikir tingkat tinggi atau sering disebut dengan HOTS dan keterampilan berpikir tingkat rendah atau disebut LOTS. Kemampuan berpikir tingkat rendah melibatkan kemampuan mengingat (CI), memahami (C2) dan menerapkan (C3) sementara dalam kemampuan berpikir tingkat tinggi melibatkan analisis (C4), mengevaluasi (C5), dan mencipta atau kreativitas (C6) Keterampilan berpikir tingkat tinggi dapat membuat seseorang individu mampu menafsirkan dan menganalisis informasi yang diperoleh (Yee dkk, 2015). Seseorang yang memiliki keterampilan berpikir tingkat tinggi tidak hanya mampu menganalisis, mengevaluasi, dan menciptakan tetapi memiliki kendali atas rencana yang dipilih, bahkan keterampilan ini membuatnya dapat beradaptasi dalam berbagai konteks (Widiawati dkk., 2018).

Pengembangan kognitif peserta didik di beberapa SMA di kota Medan untuk level HOTS masih minim diaplikasikan pada proses assessmen di sekolah. Penilaian yang mengukur kemampuan literasi kimia juga masih belum dikembangkan di sekolah tersebut. Pernyataan ini sesuai dengan penemuan pada analisis awal, yaitu sebanyak 5 sekolah sebagai responden belum menerapkan instrumen penilaian hasil belajar yang berada di kota Medan, instrumen penilaian yang digunakan masih terbatas mengukur LOTS (Lower Higher Order Thinking Skills). Menurut Ghani dkk (2017) faktor utama yang menyebabkan kemampuan berpikir peserta didik rendah yaitu karena kurangnya instrumen penilaian dan evaluasi yang efektif untuk mengukur HOTS.

Sebagian besar pendidik dalam menyusun butir soal cenderung hanya mengukur LOTS dan soal-soal yang dibuat tidak kontekstual. Soal-soal yang disusun oleh pendidik umumnya mengukur kemampuan mengingat. Butir soal yang terkandung dalam penilaian harian seperti ulangan, penugasan, dan penilaian akhir semester belum dapat sepenuhnya mengukur HOTS

peserta didik. Karena hanya mampu menyajikan informasi sampai kemampuan menerapkan ilmu yang diperoleh dari proses pembelajaran. Penilaian HOTS memiliki ciri utama yaitu mampu mengembangkan tingkat keterampilan peserta didik untuk berpikir secara kritis, kreatif, dan percaya diri (Widana, 2017).

Pendekatan yang banyak dipakai dalam analisis hasil ujian adalah pendekatan teori tes klasik (*Clasical Test Theory*). Menurut Alviah dkk, (2020) Penggunaan skor mentah/*raw score* untuk mengukur kemampuan peserta didik mempunyai kelemahan karena makna kuantitatif yang lemah. Analisis data pada penelitian ini menggunakan pemodelan Rasch. Pemodelan Rasch memiliki kelebihan mampu memprediksi data yang hilang, mampu memprediksi adanya tebakan dan mampu menganalisis kemampuan masing-masing peserta didik ditinjau dari tingkat kesulitan butir soal (Sumintono & Widhiarso, 2018).

Dalam melakukan evaluasi hasil belajar peserta didik diperlukan sebuah instrumen penilaian. Instrumen penilaian merupakan alat ukur yang memiliki fungsi dan peran vital untuk mengetahui efektivitas tahap belajar yang meliputi kemajuan output belajar peserta didik yang meliputi segi domain kognitif, afektif dan psikomotorik baik secara kelompok maupun individu (Arifin, 2009).

Berdasarkan hasil penelitian terdahulu diperoleh bahwa kemampuan HOTS peserta didik dapat meningkat dengan memanfaatkan lembar penilaian aktivitas (Ghani dkk, 2017). Penelitian lainnya melaporkan bahwa peserta didik yang diberikan lembar kerja HOTS dan bukan HOTS memeproleh hasil yang berbeda, dengan rata-rata peserta didik yang menggunakan lembar kerja HOTS lebih tinggi (Yennita, dkk, 2018). Kemudian terjadi peningkatan pada kemampuan peserta didik yang cenderung untuk berpikir lebih tinggi dengan menggunakan instrumen penilaian HOTS yang mengukur level pengetahuan (Nurhayati & Ningrum, 2016). Dalam penelitian lainnya, instrumen penilaian terdigitalisasi yang dikembangkan layak menjadi alat penilaian pembelajaran fisika di SMA dengan tingkat validitas dan reliabilitas baik (Martin, dkk, 2018).

Penelitian ini diselenggarakan dengan implikasi untuk memperoleh instrumen penilaian hasil belajar yang valid dan reliabel dalam mengukur kemampuan HOTS dan ketercapaian indikator pembelajaran oleh peserta didik. Baik pendidik maupun peneliti dapat memanfaatkan peluang ini dalam mengembangkan instrumen penilaian yang mampu mengukur HOTS. Sehingga dapat disimpulkan, instrumen penilaian merupakan bagian terpenting dalam proses pembelajaran yang memiliki tujuan mencari tahu kualitas output pembelajaran peserta didik serta mengukur keterampilan terhadap suatu materi tertentu. Kemampuan berpikir sangat penting dalam proses pendidikan. Seseorang yang berpikir dapat mempengaruhi kemampuan belajar, kecepatan dan efektivitas pembelajaran. Apabila peserta didik memiliki kemampuan berpikir tingkat tinggi maka hal itu mengembangkan diri dalam membuat keputusan, penilaian, dan menyelesaikan masalah dengan tepat.

Metode

Penelitian ini merupakan penelitian deskriptif kuantitatif. Penelitian deskriptif kuantitatif merupakan penggambaran masalah penelitian melalui deskripsi suatu keadaan atau kebutuhan akan penjelasan tentang hubungan antar variabel. Metode yang digunakan dalam penelitian ini adalah mengumpulkan data atau informasi dari para validator ahli untuk menentukan valid atau tidak valid terhadap butir soal yang dikembangkan. Partisipan penelitian ini adalah 3 validator ahli yang berasal dari dosen dan 5 validator ahli yang berasal dari guru kimia serta 70 siswa untuk mengerjakan butir soal berbasis HOTS. Pengumpulan data dilakukan dengan menggunakan kuesioner dengan jenis data kualitatif dan menggunakan *winstep* sebagai permodelan *rasch* dengan jenis data kuantitatif. Data kuantitatif merupakan data numerik yang

dikumpulkan dari sampel menggunakan instrumen pertanyaan dan respon. Data penelitian yang dihasilkan digunakan untuk menggambarkan kualitas butir tes yang dikembangkan.

Hasil

Design (Desain)

Dalam tahap ini, kisi – kisi soal dikembangkan dari indikator ketercapaian kompetensi yang telah dianalisis. Pengembangan kisi – kisi soal dimulai dengan memetakan letak dimensi pengetahuan yang akan diukur oleh butir soal sesuai dengan tabel taksonomi dalam taksonomi Bloom revisi. Hasil pemetaan dalam tabel taksonomi terdapat pada Tabel 1.

Tabel 1. Pemetaan dalam Tabel Taksonomi pada Materi Semester Genap SMA

	C1.	C2.	C3.	C4.	C5.	C6.
Faktual	-	-	-	9 soal	2 soal	-
Konspetual	-	-	-	16 soal	3 soal	1 soal
Prosedural	-	-	-	5 soal	-	1 soal
Metakognitif	-	-	-	1 soal	-	2 soal

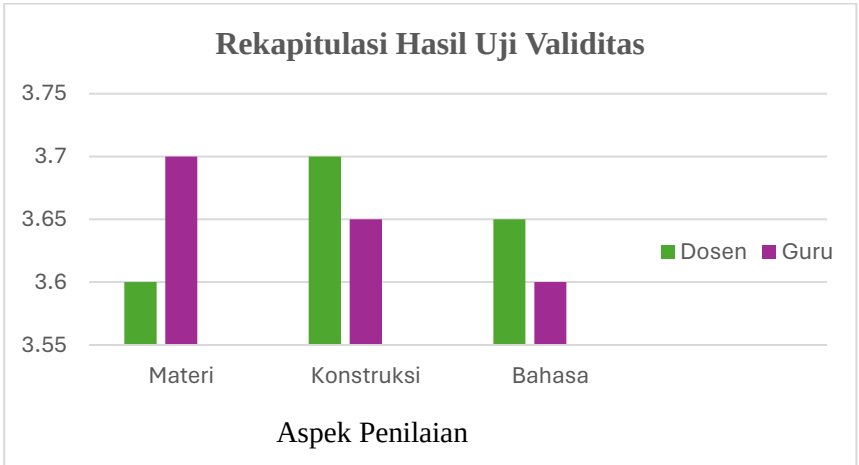
Dengan berpedoman pada Tabel 1 selanjutnya dilakukan penyusunan kisi – kisi dengan memperhatikan penggunaan kata kerja operasional (KKO) sesuai dengan dimensi yang terukur. Dan kemudian dilakukan penyusunan butir soal. Butir soal dikembangkan dalam bentuk pilihan ganda yang tersusun dari satu pilihan jawaban benar diantara lima pilihan jawaban (A, B, C, D, E) dan dikembangkan dari berbagai media cetak (bahan ajar), artikel imiah, dan sumber lainnya. Instrumen soal yang dihasilkan pada tahap ini adalah instrumen draf 1.

Development (Pengembangan)

Pada tahap ini, pengembangan instrumen draft 1 dilakukan berdasarkan hasil yang diperoleh dari beberapa pengujian. Uji yang dilakukan, yaitu uji validitasisi dan keterbacaan *one to one* dan uji validitas empiris dan analisis butir soal. Dalam tahap pengembangan dihasilkan instrumen yang siap diimplementasikan.

Uji Validitas Isi dan Keterbacaan One to One

Pertama dalam tahap pengembangan, istrumen draft 1 di uji validitas isi dan uji keterbacaan *one to one*. Instrumen yang dikembangkan divalidasi oleh beberapa ahli, yaitu 3 orang dosen kimia dan 5 orang guru kimia. Rekapitulasi hasil uji validitas isi oleh dosen dan guru dapat di lihat pada Gambar 1.



Gambar 1. Rekapitulasi Hasil Uji Validitas Isi

Berdasarkan Gambar 1 diperoleh rata – rata hasil uji validitas oleh dosendan guru *berada* pada rentang > 3,5 sehingga dikategorikan valid dan tidak perlu revisi. Hasil analisis secara kualitatif diperoleh bahwa terdapat butir soal yang perlu diperbaiki karena belum termasuk kedalam soal HOTS, seperti pada soal nomor 8 dan 11.

Tabel 2. Validasi Isi Aiken's V

Soal	Penilai			S1	S2	S3	∑S	n(c-1)	V	KET
	I	II	III							
Soal_01	4	5	4	3	4	3	10	12	0,833333	Tinggi
Soal_02	5	4	5	4	3	4	11	12	0,916667	Tinggi
Soal_03	5	5	5	4	4	4	12	12	1	Tinggi
Soal_04	4	4	5	3	3	4	10	12	0,833333	Tinggi
Soal_05	5	5	4	4	4	3	11	12	0,916667	Tinggi
Soal_06	4	5	4	3	4	3	10	12	0,833333	Tinggi
Soal_07	5	5	5	4	4	4	12	12	1	Tinggi
Soal_08	4	4	3	3	3	2	8	12	0,666667	Sedang
Soal_09	5	4	4	4	3	3	10	12	0,833333	Tinggi
Soal_10	5	5	5	4	4	4	12	12	1	Tinggi
Soal_11	4	4	3	3	3	2	8	12	0,666667	Sedang
Soal_12	5	5	5	4	4	4	12	12	1	Tinggi
Soal_13	4	4	5	3	3	4	10	12	0,833333	Tinggi
Soal_14	4	4	4	3	3	3	9	12	0,75	Sedang
Soal_15	5	5	4	4	4	3	11	12	0,916667	Tinggi
Soal_16	4	4	5	3	3	4	10	12	0,833333	Tinggi
Soal_17	4	5	4	3	4	3	10	12	0,833333	Tinggi
Soal_18	5	4	4	4	3	3	10	12	0,833333	Tinggi
Soal_19	4	4	5	3	3	4	10	12	0,833333	Tinggi
Soal_20	5	5	4	4	4	3	11	12	0,916667	Tinggi
Soal_21	4	5	5	3	4	4	11	12	0,916667	Tinggi
Soal_22	5	4	5	4	3	4	11	12	0,916667	Tinggi
Soal_23	5	5	5	4	4	4	12	12	1	Tinggi
Soal_24	4	3	5	3	2	4	9	12	0,75	Sedang
Soal_25	5	4	4	4	3	3	10	12	0,833333	Tinggi
Soal_26	5	5	4	4	4	3	11	12	0,916667	Tinggi
Soal_27	5	5	4	4	4	3	11	12	0,916667	Tinggi
Soal_28	5	5	4	4	4	3	11	12	0,916667	Tinggi
Soal_29	5	4	4	4	3	3	10	12	0,833333	Tinggi
Soal_30	5	4	5	4	3	4	11	12	0,916667	Tinggi
Soal_31	5	4	5	4	3	4	11	12	0,916667	Tinggi
Soal_32	5	4	5	4	3	4	11	12	0,916667	Tinggi
Soal_33	4	4	5	3	3	4	10	12	0,833333	Tinggi
Soal_34	4	4	5	3	3	4	10	12	0,833333	Tinggi
Soal_35	5	5	4	4	4	3	11	12	0,916667	Tinggi
Soal_36	4	5	4	3	4	3	10	12	0,833333	Tinggi
Soal_37	4	5	4	3	4	3	10	12	0,833333	Tinggi
Soal_38	4	5	4	3	4	3	10	12	0,833333	Tinggi

Soal_39	4	4	4	3	3	3	9	12	0,75	Sedang
Soal_40	5	4	3	4	3	2	9	12	0,75	Sedang
Soal_41	5	4	3	4	3	2	9	12	0,75	Sedang
Soal_42	5	4	5	4	3	4	11	12	0,916667	Tinggi
Soal_43	5	4	5	4	3	4	11	12	0,916667	Tinggi
Soal_44	5	4	5	4	3	4	11	12	0,916667	Tinggi
Soal_45	5	4	5	4	3	4	11	12	0,916667	Tinggi
TOTAL	207	199	199	162	154	154	470	540	0,87037	Tinggi
									AVERAGE	

Uji keterbacaan dilakukan untuk mengetahui komentar siswa terhadap butir soal hasil *pengembangan*. Uji keterbacaan diterapkan denganmengumpulkan 10 orang siswa non sampel dan siswa diminta untuk membaca butir soal. Kemudian siswa dipersilakan untuk menyampaikan komentar mengenai kejelasan bahasa yang digunakan dan memeriksa kesalahan pengetikan. Tanggapan siswa dirangkum dalam Tabel 3.

Tabel 3. Uji Keterbacaan *One to One*

Item	Hasil Uji Keterbacaan
1	Menambahkan kalimat “Perhatikan data berikut” pada awal soal
4	Mengganti kata “dari” menjadi “berdasarkan”
6	Menambah satuan g/mol pada data berat molekul
16	Mengganti “air bags” menjadi “air bag”
18	Memperbaiki spasi pada persamaan kimia
23	Memperbaiki penulisan rumus kimia
25	Memperbaiki penulisan kata “table”
35	Memperjelas gambar yang terdapat pada soal
40	Memperbaiki penulisan satuan pada soal

Berdasarkan Tabel 3 ditemukan kesalahan pengetikan di beberapa nomor soal. Kekurangan yang ditemukan dari hasil pengujian disempurnakan dan menghasilkan instrumen draft 2. Kemudian instrumen draft 2 di uji coba ke kelompok A sebelum di uji coba ke kelompok B. Uji coba ini dilakukan untuk melihat validitas konstruk serta validitas empiris dan kualitas butir soal yang terkategori baik dan dapat digunakan sebagai alat dalam mengukur hasil belajar.

Validitas Empiris dan Analisis Butir Soal

Setelah instrumen di uji validitas isi, kemudian instrumen di uji coba ke kelompok sampel A yang terdiri dari 70 sampel siswa. Uji coba ini dilakukan untuk memeriksa validitas item secara empiris, reliabilitas, tingkat kesukaran, daya pembeda, dan distraktor. Pengujian ini dilakukan dengan menggunakan analisis model *Rasch.* Pemodelan *Rasch* menghubungkan parameter butir dengan kemampuan peserta tes. Analisis dilakukan dengan menggunakan bantuan program komputer yaitu *Winstep*.

Validitas Empiris

Item dikatakan valid jika item cocok (*fit*) dengan model *Rasch*. Kecocokan dengan model terpenuhi jika dua kategori sesuai dengan kriteria. Validitas empiris dapat dilihat dari hasil *Outfit* MNSQ, ZSTD,dan *Pt. Measure Corr* pada Tabel 4.

Tabel 4. Hasil *Outfit* MNSQ, ZSTD, dan *Pt. Measure Corr.*

Item	Outfit		Pt. Measure Corr.	Keterangan
	MNSQ	ZSTD		
1	1,11	1,1	0,12	*Fit*
2	1,05	0,4	0,17	*Fit*
3	0,88	-1,4	0,44	*Fit*
4	1,23	1,3	0,00	*Fit*
5	0,82	-1,4	0,49	*Fit*
6	1,36	1,7	0,05	*Fit*
7	1,86	1,1	-0,4	*Misfit*
8	1,40	1,2	0,3	*Fit*
9	0,65	-2,0	0,62	*Fit*
10	0,61	-2,0	0,64	*Fit*
11	1,06	0,3	0,13	*Fit*
12	1,07	0,5	0,19	*Fit*
13	0,80	-1,6	0,53	*Fit*
14	1,14	1,0	0,16	*Fit*
15	0,48	-1,5	0,54	*Fit*
16	0,62	-1,3	0,55	*Fit*
17	1,08	1,0	0,19	*Fit*
18	0,80	-2,2	0,56	*Fit*
19	0,96	-0,4	0,33	*Fit*
20	0,95	-0,5	0,35	*Fit*
21	0,84	-1,6	0,50	*Fit*
22	0,96	-0,2	0,32	*Fit*
23	1,25	1,4	0,15	*Fit*
24	0,86	-1,3	0,48	*Fit*
25	1,45	3,7	-0,16	*Fit*
26	1,22	1,8	0,10	*Fit*
27	1,26	2,9	-0,03	*Fit*
28	1,20	2,1	0,30	*Fit*
29	1,53	2,7	-0,20	*Misfit*
30	1,07	0,4	0,29	*Fit*
31	0,85	-1,2	0,47	*Fit*
32	0,58	-2,1	0,66	*Fit*
33	0,71	-2,2	0,61	*Fit*
34	1,58	2,8	-0,2	*Misfit*
35	1,18	1,9	0,10	*Fit*
36	2,17	1,5	-0,14	*Misfit*
37	1,00	0,0	0,31	*Fit*
38	1,20	1,1	0,0	*Fit*
39	0,97	-0,2	0,31	*Fit*
40	1,00	0,0	0,28	*Fit*
41	0,81	-1,3	0,46	*Fit*
42	0,95	-0,2	0,32	*Fit*
43	0,72	-2,7	0,65	*Fit*
44	0,94	-0,6	0,38	*Fit*
45	1,03	0,4	0,28	*Fit*

Dari Tabel 4 diketahui terdapat empat item yang *misfit* (tidak cocok) dengan model *Rasch*, yaitu item nomor 7, 29, 34, dan 36 dengan nilai *Outfit* MNSQ 1,86; 1,53; 1,58; dan 2,17. Untuk mengetahui alasan item nomor 7, 29, 34, dan 36 *misfit* maka dilakukan analisis lebih lanjut pada bagian *Item: Response.* Item nomor 7, 29, 34, dan 36 *misfit* disebabkan karena adanya *outliers* pada sampel (*person*) ditandai dengan nilai *Z – Residual* ≥ 2. Angka ini menunjukkan adanya jawaban yang tak terduga (*unexpected answers*) yang diberikan oleh sampel (*person*). Adanya *outliers* dalam analisis menyebabkan hasil analisis kecocokan item terganggu sehingga menjadi kurang dapat dipercaya, oleh karena itu *outliers* perlu dihilangkan. Terdapat 20 sampel yang harus dikeluarkan dalam analisis yaitu sampel nomor 1, 2, 17, 18, 20, 28, 29, 30, 32, 33, 34, 35, 42, 51, 52, 53, 54, 59, 62, 67.

Setelah *outliers* dihilangkan, jumlah sampel yang tersisa adalah 50 sampel. Kemudian analisis *Item fit* dilakukan kembali untuk melihat kecocokan item dengan model. Hasil luaran analisis setelah *outliers* dihilangkan terdapat dalam Tabel 5. Pada Tabel 5 menunjukkan bahwa keseluruhan item cocok dengan model (*fit*). Oleh karena itu analisis reliabilitas dapat dilanjutkan *Measure Corr* pada Tabel 5.

Tabel 5. Hasil *Outfit* MNSQ, ZSTD, dan *Pt. Measure Corr* Setelah *Outliers* Dihilangkan

Item	Outfit		Pt. Measure Corr.	Keterangan
	MNSQ	ZSTD		
1	1,30	1,9	-0,20	*Fit*
2	1,21	1,2	-0,04	*Fit*
3	0,96	-0,3	0,28	*Fit*
4	1,38	1,9	-0,26	*Fit*
5	0,91	-0,3	0,28	*Fit*
6	0,92	-0,3	0,32	*Fit*
7	1,01	0,2	0,45	*Fit*
8	1,01	0,2	0,09	*Fit*
9	0,68	-1,0	0,45	*Fit*
10	0,76	-0,3	0,24	*Fit*
11	1,49	1,2	-0,13	*Fit*
12	1,04	0,3	0,23	*Fit*
13	0,75	-1,9	0,63	*Fit*
14	0,98	-0,1	0,29	*Fit*
15	0,76	-0,6	0,34	*Fit*
16	0,56	-0,6	0,34	*Fit*
17	1,08	1,0	0,11	*Fit*
18	0,90	-0,6	0,33	*Fit*
19	0,94	-0,6	0,33	*Fit*
20	1,02	0,3	0,22	*Fit*
21	0,87	-0,9	0,42	*Fit*
22	1,06	0,4	0,22	*Fit*
23	0,91	-0,5	0,35	*Fit*
24	0,69	-1,9	0,65	*Fit*
25	1,19	1,6	0,04	*Fit*
26	1,02	0,2	0,29	*Fit*
27	1,11	1,3	0,08	*Fit*
28	0,96	-0,4	0,32	*Fit*

29	1,03	0,2	0,10	*Fit*
30	0,87	-0,7	0,48	*Fit*
31	1,00	0,1	0,20	*Fit*
32	0,59	-0,6	0,34	*Fit*
33	0,78	-0,7	0,37	*Fit*
34	1,09	0,4	0,16	*Fit*
35	1,16	1,7	0,04	*Fit*
36	1,01	0,2	0,45	*Fit*
37	0,95	-0,5	0,32	*Fit*
38	1,46	1,9	-0,31	*Fit*
39	0,90	-0,7	0,39	*Fit*
40	1,23	1,9	-0,07	*Fit*
41	0,93	-0,6	0,35	*Fit*
42	0,92	-0,3	0,32	*Fit*
43	0,73	-1,3	0,54	*Fit*
44	0,84	-1,3	0,48	*Fit*
45	0,84	-1,9	0,51	*Fit*

Tabel 5 menunjukkan bahwa terjadi penurunan skala nilai pada rentangnilai *OUTFIT* MNSQ setelah *outliers* pada person dikeluarkan sehingga totalsampel 50 orang. Ketika *outliers* belum dikeluarkan (sampel = 70) item nomor 7 misfit dengan nilai 1,86. Kemudian setelah pengurangan *outliers* (sampel = 50) item nomor 7 fit dengan nilai 1,01. Telah terjadi penurunan sebanyak 0,85 akibat pengaruh pengurangan sampel yang mengganggu analisis *fit* dalam pemodelan *Rasch*. Penentuan *item fit* dapat pula dilihat dari nilai *Outfit ZSTD* dan *PT. Measure Corr.* Apabila nilai *Outfit MNSQ* masih tetap tidak memenuhi kriteria maka kecocokan item dilihat dari nilai *Outfit ZSTD* dan *PT. Measure Corr.*

Reliabilitas

Penentuan reliabilitas dilakukan untuk melihat konsistensi instrumen dalam pengukuran jika digunakan secara berulang. Nilai reliabilitas secara keseluruhan dapat dilihat dari nilai *alpha Cronbach*. Untuk mengetahui tingkat konsistensi responden dapat dilihat dari hasil *Person Reliability*.Sedangkan untuk melihat kualitas per item dalam instrumen ditentukan dari nilai *Item Reliability*. Hasil analisis nilai yang diperoleh berada pada Tabel 8.

Tabel 6. Hasil Analisis Reliabilitas

Analisis Reliabilitas	Reliabilitas	Kategori
Alpha Cronbach	0,76	Bagus
Person Reliability	0,71	Cukup
Item Reliability	0,91	Baik Sekali

Dari Tabel 6 diperoleh hasil bahwa instrumen yang dikembangkan reliabel dengan kriteria *alpha Cronbach* bagus, *Person Reliability* cukup, dan *Item Reliability* baik sekali.

Tingkat Kesukaran

Selanjutnya item yang telah teruji validitas dan relibailitasnya di uji tingkat kesukaran item. Butir tes yang memiliki tingkat kesukaran baik jika berada pada rentang -2 sampai +2. Hasil analisis tingkat kesukaran instrumenberada pada Tabel 7.

Tabel 7. Tingkat Kesukaran

Tingkat Kesukaran	Nomor Soal	Jumlah Soal (Persen)
Mudah	32, 15, 16	3 (6,7%)
Sedang	1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 14, 17, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 37, 38, 39, 40, 41, 42, 43, 44, 45	39 (86,7%)
Sulit	8, 36	2 (4,4%)
Sangat Sulit	7	1 (2,2%)

Dari Tabel 7 diperoleh informasi bahwa terdapat 1 butir soal dengan kategori sangat sulit (b > 2), 3 butir soal yang dikategorikan sebagai soal mudah karena nilai *item* estimate (b) berada dalam rentang ($-2 \leq b \leq -1$), 39 butir soal dengan kategori soal sedang ($-1 < b \leq 1$), dan 2 butir soal dalam kategori soal sulit ($1 < b \leq 2$).

Daya Pembeda

Informasi tentang siswa yang memiliki kemampuan tinggi atau rendah dilihat dari analisis daya pembeda. Daya pembeda dilakukan untuk melihat kemungkinan soal di jawab dengan benar atau salah oleh siswa. Jika soal dapat dijawab benar oleh siswa dengan kemampuan tinggi dan soal dijawab salah oleh siswa dengan kemampuan rendah, maka butir soal dianggap mampu membedakan kemampuan siswa. Analisis daya pembeda dilakukan dengan melihat nilai *Pt. Measure Corr* (Tabel 5). Hasil analisis daya pembeda terdapat pada Tabel 8.

Tabel 8. Daya Pembeda

Daya Pembeda Item	Nomor Soal	Jumlah Soal (Persen)
Sangat Bagus	7, 9, 13, 21, 24, 30, 36, 43, 44, 45	10 (22,2%)
Bagus	6, 15, 16, 18, 19, 23, 28, 32, 33, 37, 39, 41, 42	13 (28,8%)
Cukup	3, 5, 10, 12, 14, 20, 22, 26, 31	9 (20%)
Tidak Bisa Mendiskriminasi	8, 17, 25, 27, 29, 34, 35	7 (15,5%)
Membutuhkan Pemeriksaan Terhadap Butir (Buruk)	1, 2, 4, 11, 38, 40	6 (13,5%)

Berdasarkan Tabel 8 diperoleh hasil bahwa terdapat 13 dari 45 butir soal yang daya pembedanya kurang berfungsi dengan baik karena nilai *Pt. Measure Corr* berada pada rentang < 0,20. Dan terdapat 32 butir soal yang berada pada rentang ≥ 0,20. Hasil ini menunjukkan bahwa butir soal memiliki daya pembeda yang dapat berfungsi dengan baik.

Distraktor

Analisis distraktor dalam pemodelan *Rasch* dilihat dari hasil rata-rata nilai *logit*. Hasil pengujian distraktor dengan menggunakan program *Winsteps* terdapat pada Tabel 9.

Tabel 9. Distraktor

Pengecoh	Nomor Soal	Jumlah Soal (Persen)
Berfungsi Baik	3, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 24, 26, 28, 29, 30, 31, 32, 33, 34, 36, 37, 39, 41, 42, 43, 44, 45	35 (77%)
Tidak Berfungsi	1, 2, 4, 11, 17, 25, 27, 35, 38, 40	10 (23%)
	Jumlah	45 (100 %)

Analisis distraktor pada Tabel 9 diperoleh sebanyak 35 butir soal yang berfungsi dengan baik dengan presentase 77% dan sebanyak 10 butir soal dengan persentase 23% dikategorikan tidak berfungsi.

Dari hasil analisis butir soal diperoleh jumlah butir soal yang layak sebanyak 32 butir dan sebanyak 13 butir soal dikategorikan tidak layak digunakan sebagai alat untuk mengukur hasil belajar. Butir soal dikategorikan tidak layak karena tidak memenuhi satu dari lima analisis butir soal. Data 32 butir soal yang layak dari total keseluruhan 45 soal disajikan pada Tabel 10.

Tabel 10 Ringkasan Hasil Analisis Butir Soal

No Soal	Validitas	Reliabilitas	Tingkat Kesukaran	Daya Pembeda	Distraktor	Ket
1	Valid	Reliabel	Sedang	**Buruk**	**Tidak Baik**	**Tolak**
2	Valid	Reliabel	Sedang	**Buruk**	**Tidak Baik**	**Tolak**
3	Valid	Reliabel	Sedang	Cukup	Baik	Terima
4	Valid	Reliabel	Sedang	**Buruk**	**Tidak Baik**	**Tolak**
5	Valid	Reliabel	Sedang	Cukup	Baik	Terima
6	Valid	Reliabel	Sedang	Bagus	Baik	Terima
7	Valid	Reliabel	Sangat Sulit	Sangat Bagus	Baik	Terima
8	Valid	Reliabel	Sulit	**Tidak Mampu Mendiskriminasi**	Baik	**Tolak**
9	Valid	Reliabel	Sedang	Sangat Bagus	Baik	Terima
10	Valid	Reliabel	Sedang	Cukup	Baik	Terima
11	Valid	Reliabel	Sedang	**Buruk**	**Tidak Baik**	**Tolak**
12	Valid	Reliabel	Sedang	Cukup	Baik	Terima
13	Valid	Reliabel	Sedang	Sangat	Baik	Terim

				Bagus		a
14	Valid	Reliabel	Sedang	Cukup	Baik	Terima
15	Valid	Reliabel	Mudah	Bagus	Baik	Terima
16	Valid	Reliabel	Mudah	Bagus	Baik	Terima
17	Valid	Reliabel	Sedang	**Tidak Mampu Mendiskriminasi**	**Tidak Baik**	**Tolak**
18	Valid	Reliabel	Sedang	Bagus	Baik	Terima
19	Valid	Reliabel	Sedang	Bagus	Baik	Terima
20	Valid	Reliabel	Sedang	Cukup	Baik	Terima
21	Valid	Reliabel	Sedang	Sangat Bagus	Baik	Terima
22	Valid	Reliabel	Sedang	Cukup	Baik	Terima
23	Valid	Reliabel	Sedang	Bagus	Baik	Terima
24	Valid	Reliabel	Sedang	Sangat Bagus	Baik	Terima
25	Valid	Reliabel	Sedang	**Tidak Mampu Mendiskriminasi**	**Tidak Baik**	**Tolak**
26	Valid	Reliabel	Sedang	Cukup	Baik	Terima
27	Valid	Reliabel	Sedang	**Tidak Mampu Mendiskriminasi**	**Tidak Baik**	**Tolak**
28	Valid	Reliabel	Sedang	Bagus	Baik	Terima
29	Valid	Reliabel	Sedang	**Tidak Mampu Mendiskriminasi**	Baik	**Tolak**
30	Valid	Reliabel	Sedang	Sangat Bagus	Baik	Terima
31	Valid	Reliabel	Sedang	Cukup	Baik	Terima
32	Valid	Reliabel	Mudah	Bagus	Baik	Terima
33	Valid	Reliabel	Sedang	Bagus	Baik	Terim

						a
34	Valid	Reliabel	Sedang	**Tidak Mampu Mendiskriminasi**	Baik	**Tolak**
35	Valid	Reliabel	Sedang	**Tidak Mampu Mendiskriminasi**	**Tidak Baik**	**Tolak**
36	Valid	Reliabel	Sulit	Sangat Bagus	Baik	Terima
37	Valid	Reliabel	Sedang	Bagus	Baik	Terima
38	Valid	Reliabel	Sedang	**Buruk**	**Tidak Baik**	**Tolak**
39	Valid	Reliabel	Sedang	Bagus	Baik	Terima
40	Valid	Reliabel	Sedang	**Buruk**	**Tidak Baik**	**Tolak**
41	Valid	Reliabel	Sedang	Bagus	Baik	Terima
42	Valid	Reliabel	Sedang	Bagus	Baik	Terima
43	Valid	Reliabel	Sedang	Sangat Bagus	Baik	Terima
44	Valid	Reliabel	Sedang	Sangat Bagus	Baik	Terima
45	Valid	Reliabel	Sedang	Sangat Bagus	Baik	Terima

Selanjutnya dari 32 butir soal yang dikategorikan layak dipilih 20 soal yang dianggap mampu mempresentasikan setiap kompetensi dasar dari indikator pencapaian kompetensi. Soal yang dihasilkan dari hasil analisis butir soal dianggap sebagai soal draft 3.

Pembahasan

Dalam tahap *design* dilakukan pengembangan kisi – kisi soal. Pengembangan kisi – kisi soal dimulai dengan memetakan letak dimensi pengetahuan sesuai dengan tabel taksonomi dalam taksonomi Bloom revisi. Tabel taksonomi merupakan gabungan antara dimensi pengetahuan dan dimensi proses kognitif. Pemetaan ini dilakukan agar penyusunan butir soal lebih terstruktur.

Selanjutnya dilakukan penyusunan butir soal. Jumlah butir soal yang dikembangkan dari 3 kompetensi dasar sebanyak 45 soal. Penyusunan butir soal HOTS dilengkapi dengan stimulus yang dijadikan sebagai rujukan pertanyaan. Penyajian stimulus bersumber dari keadaan terkini yang merupakan fakta di lingkungan sekitar daerah pendidikan (kontekstual) dan menarik (Suhardjanto, 2021). Penilaian yang berorientasi HOTS membantu guru dalam mengevaluasi proses pembelajaran yang tepat untuk pemebelajaran berikutnya dan membantu siswa meningkatkan kemampuan berpikir dan kinerja (Putri dkk., 2021).

Butir soal yang telah dikembangkan di validasi ke validator ahli, yaitu dosen dan guru dengan menggunakan instrumen validasi isi dan di uji keterbacaan one to one oleh 10 orang siswa. Validasi isi diperlukan ketika mengembangkan suatu produk dan informasi yang

diperoleh dari hasil validasi dapat dijadikan sebagai bahan acuan melakukan revisi (Shi dkk., 2021, Padilla & Akers, 2020). Hasil rata – rata yang diperoleh dari dosen dan guru berturut – turut sebesar 3,74 dan 3,80. Hasil ini menunjukkan bahwa instrumen yang dikembangkan valid dan tidak perlu revisi. Hasil validitas isi menunjukkan bahwa instrumen yang dikembangkan layak digunakan sebagai alat untuk mengukur HOTS siswa. Validitas isi pada instrumen dilakukan untuk mengetahui tingkat kelayakan instrumen sebelum digunakan dilapangan (Savira dkk., 2019).

Meskipun hasil validitas isi berada pada kategori baik, namun terdapat item yang masih belum termasuk soal HOTS yaitu item nomor 34 dan 38. Item nomor 34 dan 38 hanya mengukur kemampuan (C3) dan termasuk ke dalam LOTS. Oleh karena itu, diperlukan perbaikan pada stimulus dan penyusunan kalimat pertanyaan pada item nomor 34 dan 38 agar menjadi soal HOTS. Perbaikan pada item perlu dilakukan agar item yang dikembangan sesuai dengan indikator dan level berpikir HOTS (Sappaile & Pristiwaluyo, 2019). Sedangkan uji keterbacaan one to one diperoleh beberapa kesalahan pengetikan kata pada butir soal dan perlu di revisi. Komentar dan kritik dari tahap one to one diperbaiki dan instrumen penilaian HOTS siap untuk di uji coba secara empiris (Saputro dkk., 2019).

Butir soal selanjutnya di uji validitas empiris, reliabilitas, tingkat kesukaran, daya pembeda, dan distraktor. Terdapat hasil luaran Infit dan Outfit dalam pemodelan Rasch. Penentuan validitas empiris dilihat dari kecocokan (fit) dengan model. Secara umum hasil luaran outfit lebih baik digunakan dalam penentuan suatu item dikategorikan fit karena lebih peka terhadap outliers, yaitu data yang mengganggu hasil pengukuran dalam analisis (Linacre, 2012).

Pada Tabel 4.9 dari 40 item yang dikembangkan terdapat 4 item, yaitu item nomor 7, 29, 34, dan 36 yang tidak cocok dengan model (misfit) dengan nilai outfit MNSQ sebesar 1,86; 1,53; 1,58; dan 2,17. Oleh karena dibutuhkan proses identifikasi hubungan antara item dan person pada statistik Item: Response. Hasil statistik menunjukkan sampel nomor 1, 2, 17, 18, 20, 28, 29, 30, 32, 33, 34, 35, 42, 51, 52, 53, 54, 59, 62, 67 memiliki nilai Z – Residual > 2 yang dikategorikan memiliki jawaban tak terduga (unexpected answers) hal ini membuktikan bahwa terdapat sebagian siswa yang menjawab dengan cara menebak jawaban. Jawaban dari sampel ini perlu dihilangkan karena mempengaruhi item yang misfit (Boone dkk., 2013). Setelah dilakukan analisis item fit ulang diperoleh item nomor 7, 29, 34, dan 36 fit dengan model dan memiliki nilai outfit MNSQ 1,01; 1,03; 1,09; dan 1,01. Jumlah total sampel adalah 50 orang.

Pada hasil analisis outfit ZSTD terdapat beberapa item yang memiliki nilai lebih kecil dari – 2 atau lebih besar +2. Hasil ini tidak menjadi poin utama dalam penentuan kecocokan dengan model. Analisis lanjutan dengan yang merujuk pada nilai ZSTD dilakukan jika setelah penghapusan outliers nilai MNSQ tetap tidak terpenuhi (Boone dkk., 2013).

Setelah item terbukti cocok dengan model, dilakukan analisis reliabilitas. Reliabilitas menjadi penentu hasil pengukuran yang dilakukan secara repetisi pada item jika item digunakan pada sampel (person) lain, ataupun sebagai penentu hasil pengukuran sampel (person) jika digunakan pada keadaan lain (Aryadoust dkk., 2021). Pada Tabel 6 reliabilitas item yang diperoleh 0,91 dalam kategori baik sekali, reliabilitas person diperoleh sebesar 0,71 dengan kategori cukup, dan interaksi reliabilitas antar item – person sebesar 0,76 diperoleh dari nilai chronbach's alpha. Interaksi yang ditunjukkan oleh nilai chronbach's alpha antara sampel dengan item berada pada kategori diterima (Solihatun dkk., 2019).

Hasil luaran analisis tingkat kesukaran dapat dilihat dari hasil pengukuran measure yang diperoleh dari fungsi logit perbandingan jawaban benar dengan jumlah soal keseluruhan (Pratama, 2020). Tabel 7 menunjukkan bahwa item nomor 7 merupakan soal dengan kategori

sangat sulit sehingga tidak baik untuk digunakan dalam penilaian. Item yang dikategorikan sulit adalah item nomor 8 dan 36 memiliki nilai logit berturut-turut sebesar 1,15 dan 1,52. Kemudian banyak siswa yang mampu menjawab item nomor 7, 8, dan 36 berturut-turut sebanyak 2, 3, dan 4 orang.

Sedangkan untuk item dengan tingkat kesulitan yang rendah berada pada item nomor 32, 15, dan 16. Jumlah nilai logit item nomor 32, 15, dan 16 secara berturut adalah –1,19; –1,22; dan –1,45. Item lainnya adalah item yang terkategori memiliki tingkat kesulitan yang sedang dan terletak pada rentang 0,92 sampai –0,10. Kategori tingkat kesulitan item ditentukan dari jumlah siswa menjawab benar butir soal. Nilai logit dapat pula dilihat dengan menggunakan peta persebaran hubungan person – item pada Wright map. Peta tersebut digambarkan dari penggabungan nilai logit pada item dan person measure. Di bagian kiri wright map menunjukkan nilai logit person dan bagian kanan menunjukkan nilai logit item (Hamdu dkk., 2020).

Sampel 034 adalah sampel yang memiliki kemampuan paling tinggi dengan nilai logit 1,34. Namun kemampuan sampel 034 masih belum mampu menjawab item 7 dengan nilai logit 3,38 tetapi mampu menjawab item 8 dan 36 yang termasuk soal sulit. Item nomor 16 merupakan item yang mudah dengan logit –1,45. Oleh karena itu, item sulit memiliki sedikit kemungkinan di jawab benar oleh siswa dengan kemampuan rendah sedangkan kemungkinan item mudah dijawab benar oleh siswa dengan kemampuan rendah semakin besar (Alfarisa & Purnama, 2019). Item dengan kategori sedang cenderung mudah yang memiliki nilai logit mulai dari 0,94 sampai –0,95 mampu dijawab benar oleh mayoritas sampel. Namun, sampel dengan nilai logit < –0,31 tidak mampu menjawab dengan benar item 16. Hal ini mengindikasikan bahwa sampel tersebut memiliki kemampuan yang cenderung rendah.

Analisis selanjutnya adalah penentuan daya pembeda. Pada Tabel 8 diperoleh sebanyak 13 item yang tidak mampu mendiskriminasi, yaitu item nomor 1, 2, 4, 8, 11, 17, 25, 27, 29, 34, 35, 38, 40. Item ini memiliki nilai Pt. Measure Corr < 0,20 sehingga item perlu dihilangkan atau dapat pula di revisi sebelum digunakan di kelas sesungguhnya (Alagumalai dkk., 2005). Dalam penelitian ini, karena masih terdapat item yang merepresentasikan kompetensi dasar, maka item yang tidak mampu mendiskriminasi ditolak. Item yang memiliki daya pembeda dengan kategori sangat baik terdapat 22,2% serta item dengan kategori baik dan cukup sebanyak 28,8% dan 20%.

Analisis distraktor pada Tabel 9 diperoleh sebanyak 35 butir soal yang berfungsi dengan baik dengan presentasi 77% dan sebanyak 10 butir soal dengan persentasi 23% dikategorikan tidak berfungsi. Setelah dilakukan analisis kualitas butir soal maka menghasilkan produk akhir berupa butir soal yang terkategori baik untuk mengukur HOTS. Dari 45 butir yang dikembangkan, diperoleh 32 butir soal yang layak digunakan pada tahap berikutnya. Butir soal dikategorikan tidak layak karena tidak memenuhi satu dari lima analisis butir soal. Data 32 butir soal yang layak dari total keseluruhan 45 soal disajikan pada Tabel 10. Instrumen penilaian yang memiliki kriteria baik pada validitas isi, validitas empiris, reliabilitas, tingkat kesukaran, daya pembeda, dan distraktor siap digunakan sebagai instrumen untuk mengukur HOTS siswa (Utama dkk., 2020).

Conclusion

Instrumen penilaian hasil belajar kimia semester genap kelas X yang dikembangkan dikategorikan layak digunakan sebagai instrumen penilaian hasil belajar berdasarkan validasi isi dengan nilai rata-rata dari dosen dan guru sebesar 3,74 dan 3,80.

Hasil analisis item menggunakan permodelan Rasch dihasilkan 45 item fit dengan

model, reliabilitas instrumen berada pada kategori baik dengan nilai 0,71, tingkat kesukaran item yang mendominasi terkategori sedang dengan perolehan 86,7%, daya pembeda yang mampu mendiskriminasi sebanyak 71%, dan pengecoh 77% berfungsi dengan baik. Sehingga dihasilkan 32 dari 45 soal layak dan terkategori sebagai soal yang baik.

Bagi guru SMA/MA, instrumen penilaian HOTS pada materi kimia semester genap kelas X dapat digunakan sebagai instrumen mengukur kemampuan hasil belajar di sekolah dan dapat pula dijadikan pedoman penyusunan maupun analisis soal dengan pemodelan Rasch. Bagi sekolah, mempersiapkan fasilitas pendukung masih kurang ketika proses pembelajaran berlangsung dan lebih mampu untuk memanfaatkan fasilitas dan sumber daya manusia (guru) untuk memaksimalkan prestasi belajar siswa.

References

Alagumalai, S., Curtis, D. D., & Hungi, N. (2005). *Applied Rasch Measurement: A Book of Exemplars.* Dordrecht, the Netherlands: Springer.

Alfarisa, F., & Purnama, D. N. (2019). Analisis Butir Soal Ulangan Akhir Semester Mata Pelajaran Ekonomi SMA Menggunakan RASCH Model. *Jurnal Pendidikan Ekonomi Undiksha*, *11*(2), 366-374.

Alfarisa, F., & Purnama, D. N. (2019). Analisis Butir Soal Ulangan Akhir Semester Mata Pelajaran Ekonomi SMA Menggunakan RASCH Model. *Jurnal Pendidikan Ekonomi Undiksha*, *11*(2), 366-374.

Alviah, I., Susilowati, E., & Masykuri, M. (2020). Pengaruh Kemampuan Literasi Kimia Terhadap Capaian *Higher Order Thinking Skills* (HOTS) Siswa SMA Negeri 1 Sukoharjo Pada Materi Larutan Penyangga Dengan Pemodelan *Rasch*. *Jurnal Pendidikan Kimia*. Vol. 9. No. 2. 121-130.

Arifin, Z. (2009). *Evaluasi Pembelajaran Prinsip, Teknik dan Prosedur*. Bandung: PT Remaja Rosdakarya.

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.

Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch analysis in the human sciences*. Springer Science & Business Media.

Budiman, A., & Jailani, J. (2014). Pengembangan instrumen Asesmen Higher Order Thinking Skill (HOTS) pada Mata Pelajaran Matematika SMP kelas VIII Semester 1. *Jurnal Riset Pendidikan Matematika*, 1(2), 139-15.

Chairani, R & Nurfajriani. (2020). Pengembangan *E-Modul* Berbasis *Creative Problem Solving* (CPS) Pada Materi Ikatan Kimia Kelas X IPA SMA. *Jurnal Guru Kita*. 6(4).

Ghani, I. A., Ibrahim, N. H., Yahaya, N. A., & Surif, J. (2017). Enhancing students' HOTS in laboratory educational activity by using concept map as an alternative assessment tool. *Chemistry education research and practice*, *18*(4), 849-874.

Hamdu, G., Fuadi, F. N., Yulianto, A., & Akhirani, Y. S. (2020). Items Quality Analysis Using Rasch Model To Measure Elementary School Students' Critical Thinking Skill On Stem Learning. *JPI (Jurnal Pendidikan Indonesia)*, *9*(1), 61-74.

Hanifah, Nurdinah. (2019). Pengembangan Intrumen Penilaian Higher Order Thinking Skill (HOTS) disekolah dasar. Current Research in Education: *Conference Series Journal.*Vol. 1. No. 1.

Linacre, J. M. (2012). *A User's Guide to Winstep-Ministep Rasch-Model Computer Programs.*: Program Manual 3.73. 0. 2011

Martin, M., Supriyati, Y., & Budi, A. S. (2018). Pengembangan Computer Based Test (CBT) sebagai alat penilaian pembelajaran fisika SMA pada materi gerak lurus. In *Quantum: Seminar Nasional Fisika, dan Pendidikan Fisika*.

Muchlis, I, P & Andromeda. (2020). Pengembangan Instrumen Tes Berbasis *Higher Order Thinking Skill* Pada Materi Hidrolisis Garam Untuk Siswa SMA/MA. *Jurnal Eksakta Pendidikan*, Vol. 4. No. 2. 218-225.

Nurfajriani, Wildayani, H., & Nugraha, A. W. (2021). *Pengembangan Bahan Ajar Inovatif dan Interaktif Berbasis Konseptual Pada Materi Termokimia di SMA/MA*. Prosiding Seminar Nasional Kimia dan Terapan (2021). 44-49.

Nurhayati, A., & Ningrum, R. T. L. (2016). Influence of Cognitive Assessment Instrument Based Higher Order Thinking Skill Toward Students' Critical Thinking Skill. *International Conference on Mathematics, Science, and Education.*

Nurwahidah, I. (2018). Pengembangan soal penalaran model TIMSS untuk mengukur high order thinking (HOT). *THABIEA: JOURNAL OF NATURAL SCIENCE TEACHING*, *1*(1).

OECD. 2019*. PISA 2018: Assesment and Analytical Frame Work: Science, Reading, Mathematic and Financial Literacy*. Paris: OECD Publishing.

Padilla, K. L., & Akers, J. S. (2021). Content Validity Evidence for the Verbal Behavior Milestones Assessment and Placement Program. *Journal of Autism and Developmental Disorders*, 1-13.

Pratama, D. (2020). Analisis Kualitas Tes Buatan Guru Melalui Pendekatan Item Response Theory (IRT) Model Rasch. *Tarbawy: Jurnal Pendidikan Islam*, *7*(1).

Putri, C. A., Rofiqoh, E., Wulandari, F. A., Prastiningrum, F. A., & Eva, N. (2021). Asesmen Autentik: Pengembangan Asesmen HOTS Mata Pelajaran Matematika pada Siswa SMP. In *Seminar Nasional Psikologi UM*, *1*(1).

Sappaile, B. I., & Pristiwaluyo, T. (2019). Analisis butir soal ujian sekolah berstandar nasional dengan pendekatan klasik dan teori respon butir mata pelajaran matematika. In *Seminar Nasional LP2M UNM*.

Saputro, S. D., Nadliroh, N., Sari A. K., Ningsih, P. R., Wijaya, E. Y. (2019). Pengembangan Instrumen Penilaian Berbasis Hots (High Order Thinking Skill) Mata Pelajaran Sistem Komputer Kelas X SMK NEGERI 2 BANGKALAN. In *Seminar Nasional Pendidikan dan Pembelajaran 2019*

Savira, I., Wardani, S., Harjito, H., & Noorhayati, A. (2019). Desain Instrumen Tes Three Tiers Multiple Choice Untuk Analisis Miskonsepsi Siswa Terkait Larutan Penyangga. *Jurnal Inovasi Pendidikan Kimia*, *13*(1).

Shi, L., Granlund, M., Zhao, Y., Hwang, A. W., Kang, L. J., & Huus, K. (2021). Transcultural adaptation, content validity and reliability of the instrument 'Picture My Participation'for children and youth with and without intellectual disabilities in mainland China. *Scandinavian Journal of Occupational Therapy*, *28*(2), 147-157.

Solihatun, S., Rangka, I. B., Ratnasari, D., Radyati, A., Siregar, Y., Wulansari, L., & Rahim, R. (2019). Measuring of student learning performance based on geometry test for middle class in elementary school using dichotomous Rasch analysis. In *Journal of Physics: Conference Series 1157*(3).

Suhardjanto, S. (2021). Upaya Peningkatan Kemampuan Guru Bahasa Indonesia Dalam Menyusun Soal HOTS Melalui Workshop. *Jurnal Ilmiah Pro Guru*, *4*(4), 506-514.

Sumintono, B., dan Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. Cimahi: Penerbit Trim Kominikata Publishing House

Utama, C., & Nurkamto, J. (2020). The Instrument Development to Measure Higher-Order Thinking Skills for Pre-Service Biology Teacher. *International Journal of Instruction*, *13*(4).

Widana, I, W. (2017). Higher Order Thinking Skills Assessment (HOTS). *Journal of Indonesian Student Assesment and Evaluation*. Vol. 3, No. 1.

Widiawati, L., Joyoatmojo, S., & Sudiyanto, S. (2018). Higher Order Thinking Skills as Effect of Problem Based Learning in the 21st Century Learning. *International Journal of Multicultural and Multireligious Understanding*, 5(3), 96-105.

Yee, M. H., Yunos, J. M., Othman, W., Hassan, R., Tee, T. K., & Mohamad, M. M. (2015). Disparity of Learning Styles and Higher Order Thinking Skills among Technical Students. *Procedia-Social and Behavioral Sciences*, 204, 143-152.

Yennita, Y., Khasyyatillah, I., Gibran, G., & Irianti, M. (2018). Development of worksheet based on high-order thinking skills to improve high-order thinking skills of the students. *Journal of Educational Sciences*, *2*(1), 37-45.

Yudha, S., Nurfajriani., & Silaban, R. 2023. Analisis Kebutuhan Guru Terhadap Pengembangan Media Pembelajaran Kimia Berbasis Android. *Jurnal Warta Desa*. 5(1).