Jurnal Pendidikan MIPA
Volume 14. Nomor 4, Desember 2024 | ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v14i4.2063

Karakteristik Instrumen Tes *Higher Order Thinking Skills* (HOTS) Kimia SMA

Kriswantoro[1),*], **Lucya Wulandari**[2)]

[1)]Universitas Jambi
[2)]Universitas Negeri Padang

*Coresponding Author: kriswantoro18@unja.ac.id

ABSTRAK

Tujuan dari penelitian ini adalah untuk menguji alat ukur instrumen yang digunakan untuk menilai keterampilan berpikir tingkat tinggi (HOTS) siswa dalam kimia. Jenis penelitian ini yang digunakan aadalah deskriptif kuantitatif dengan menggunakan model pengembangan tes Wilson, Oriondo, dan Antonio yang terdiri atas desain tes, uji coba tes, dan pengukuran tes diterapkan dalam penyusunan ujian kimia HOTS. Instrumen pengumpulan data berupa tes HOTS kimia berbentuk multiple choice beralasan. Temuan penelitian menunjukkan bahwa tes tersebut telah terbukti sesuai dengan PCM pada tahap uji coba dan pengukuran, menurut kriteria INFIT MNSQ. Ketiga puluh elemen tersebut sesuai dengan model, menurut batas terendah dan maksimum INFIT MNSQ, yaitu 0,77 dan 1,3. Tingkat kesulitan item berkisar dari -0,88 hingga 1,04. Tes tahap uji coba memiliki peringkat reliabilitas 0,97. Tes dengan kesalahan pengukuran terkecil, 0,21, akan menghasilkan informasi terbanyak, 20,92.

Kata Kunci: Penilaian; *Higher Order Thinking Skills*

Received: 23 Okt 2024; **Revised**: 6 Nov 2024; **Accepted**: 12 Nov 2024; **Available Online**: 13 Nov 2024

PENDAHULUAN

Proses pembelajaran yang berlangsung di sekolah dan proses penilaian yang melibatkan evaluasi hasil pembelajaran terkait erat dengan kualitas pendidikan. Salah satu komponen sistem yang menentukan keberhasilan pendidikan adalah pengukuran hasil pembelajaran. Metode pengukuran yang efektif dapat dikembangkan untuk meningkatkan standar pendidikan. Proses pengukuran kegiatan harus metodis, sangat bertanggung jawab dalam pelaksanaannya, dan hasil yang diharapkan harus dapat memperhitungkan keterampilan siswa yang sebenarnya. Untuk menyederhanakan pengukuran, pelajaran kimia yang diajarkan harus mempertimbangkan kualitas unik siswa (Sunyono; et al., 2009). Mengukur pemahaman dan kemahiran siswa dalam suatu topik, serta kapasitas mereka untuk menerapkan, mensintesis, dan menganalisis konten yang diberikan oleh guru Kimia sekolah menengah, dapat menjadi cara mudah bagi para pendidik untuk mengevaluasi pembelajaran mereka.

Pembelajaran kimia tidak hanya bertujuan untuk menciptakan orang yang memahami prinsip-prinsip kimia, tetapi juga membekali peserta didik dengan pengetahuan dan kemampuan untuk menemukan dan menggunakan konsep-konsep ini dalam kehidupan sehari-hari (Kriswantoro; et al., 2021). Pembelajaran kimia tidak hanya berfokus pada teori, tetapi juga pada aplikasi dan konteks yang nyata. Selain itu, pembelajaran kimia juga harus memperkuat aspek-aspek penting seperti konten, konteks aplikasi, proses, dan sikap sains. Dengan demikian, siswa dapat memiliki pemahaman yang lebih mendalam dan relevan. Pengembangan standar moral yang tinggi dan watak ilmiah seperti objektivitas, berpikir kritis, independensi, kreativitas, dan penemuan, serta saling membantu dan keberagaman di seluruh dunia, juga termasuk dalam tujuan ini. Oleh karena itu, pembelajaran kimia di SMA tidak hanya membantu siswa memahami konsep kimia saja. Selain itu, siswa dapat mengembangkan kemampuan analitis mereka. Hal ini dilakukan agar peserta didik mampu menghadapi tantangan abad 21 yang menuntut keterampilan berpikir kritis dan pemecahan masalah. Namun, implementasi penilaian yang dapat mengukur kemampuan tersebut di bidang kimia masih kurang optimal. Keterampilan berpikir kritis ditunjukkan melalui analisis, evaluasi, deduksi, induksi, pemecahan masalah,

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v14i4.2063

mempertimbangkan berbagai pilihan, dan menarik kesimpulan (Putri & Syolendra, 2024). Selain kemampuan berpikir tingkat rendah, siswa juga memerlukan kemampuan berpikir tingkat tinggi.

Pemerintah mengharapkan para peserta didik mencapai berbagai kompetensi dengan penerapan HOTS atau keterampilan bepikir tingkat tinggi. *Higher Order Thinking Skills* (HOTS) adalah kemampuan berpikir tingkat tinggi yang dapat mendorong seseorang untuk berpikir secara luas dan mendalam tentang suatu masalah. Mengembangkan kemampuan tingkat tinggi merupakan salah satu tujuan utama pendidikan di abad ke-21. Proses berpikir yang rumit dalam mendefinisikan materi, menarik kesimpulan, membuat representasi, menganalisis, dan membentuk asosiasi yang melibatkan operasi otak yang paling mendasar merupakan salah satu deskripsi keterampilan berpikir tingkat tinggi (Resnick, 1987). Sesuai dengan tangga taksonomi Bloom, kemampuan ini juga digunakan untuk menyoroti proses berpikir tingkat tinggi. Alat ukur yang dikenal sebagai pertanyaan HOTS digunakan untuk menilai kemampuan berpikir tingkat tinggi, yaitu berpikir di luar ingatan sederhana, mengulang, atau membacakan informasi tanpa pemrosesan lebih lanjut. (Pratiwi, 2022). Namun, HOTS yang dimaksud adalah menekankan lebih banyak pada aspek menganalisis, mengevaluasi, dan mencipta (Isnawati et al., 2024). Keunggulan taksonomi Bloom yang dianggap lebih unggul dibanding taksonomi lain, pembuatan soal HOTS di Indonesia selama ini lebih banyak mengandalkan metodologi ini. Sebagaimana penelitian yang telah dilakukan untuk membuat instrumen tes HOT dengan Taksonomi Bloom (Istiyono et al., 2014). Maka dari itu suatu tes harus berkualitas. Pada soal HOTS merupakan tipe soal yang lebih kompleks dimana menawarkan beberapa solusi dan menuntut pemikiran kritis dalam proses menyelesaikannya (Faradisa et al., 2024).

Kualitas sebuah tes sangat banyak aspek yang bisa mempengaruhi. Salah satu faktor yang dipertimbangkan adalah seberapa pentingnya menilai kualitas tes berdasarkan kemampuan guru dalam melaksanakan penilaian dan membuat soal tes (Johar, 2012). Kemampuan instruktur dalam merumuskan pertanyaan sangat krusial dalam meningkatkan standar pengajaran karena kemampuan ini mendorong perkembangan kemampuan berpikir, khususnya kemampuan berpikir yang canggih (Sunggingwati & Nguyen, 2013). Pengamatan para peneliti menunjukkan bahwa sebagian besar soal ujian yang digunakan di sekolah dirancang untuk menyoroti keterampilan belajar siswa yang lemah. Siswa hanya menyerap materi secara pasif saat mengikuti ujian. Siswa tidak memperoleh pengalaman yang diperlukan untuk membangun kemampuan berpikir kritis dalam jenis pembelajaran ini, khususnya kemampuan untuk menarik kesimpulan dan memberikan penjelasan yang jelas.

Pemeriksaan soal-soal ujian tengah semester dan akhir semester menunjukkan bahwa sebagian besar soal yang dibuat guru masih tergolong dalam kelompok berpikir tingkat rendah. Ini adalah masalah penting lainnya (Iskandar & Senam, 2015). Temuan penelitian menunjukkan bahwa siswa jarang diberikan tes keterampilan berpikir tingkat tinggi (HOTS). Pemeriksaan butir-butir pertanyaan diperlukan untuk menguji setiap pertanyaan yang pada akhirnya akan digunakan untuk melaksanakan tes (Amelia & Kriswantoro, 2017). Guru perlu melakukan tugas analisis butir soal untuk meningkatkan mutu butir soal ujian tertulis. Pada akhirnya, hasil analisis akan menggambarkan fitur instrumen ujian itu sendiri.

. Hasil dari tes HOTS kimia yang telah dikembangkan perlu di analisis dan dikalibrasi untuk melihat karakteristiknya. Hal ini dilakukan sesuai dengan tujuan penelitian yaitu untuk memperoleh instrumen tes yang teruji karakteristiknya, dan akan menghasilkan perangkat tes yang akurat, valid, dan reliabel, sehingga kesimpulan yang diambil dapat lebih akurat. Hal inilah yang menjadi panduan untuk kedepannya dalam hal penilaian hasil belajar berpikir tingkat tinggi. Instrumen yang telah teruji karakteristinya akan lebih mudah dalam mengukur kemampuan peserta didik.

Penilaian tradisional yang banyak digunakan di sekolah cenderung berfokus pada keterampilan berpikir tingkat rendah, seperti mengingat dan memahami. Penilaian jenis ini belum cukup memadai untuk menggali keterampilan HOTS peserta didik. Oleh karena itu, diperlukan penelitian yang mengembangkan dan menguji instrumen penilaian HOTS yang valid, andal, dan sesuai dengan karakteristik kimia sebagai bidang ilmu yang bersifat abstrak dan analitis.vDengan adanya penelitian dan pengembangan penilaian berbasis HOTS, guru diharapkan dapat lebih kreatif dalam merancang pembelajaran kimia yang tidak hanya mengandalkan hafalan tetapi juga menantang peserta didik untuk berpikir lebih dalam. Hal ini akan meningkatkan kualitas pembelajaran dan motivasi siswa dalam belajar kimia.

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v14i4.2063

METODE

Penelitian ini menggunakan teknik deskriptif kuantitatif dan tergolong penelitian pengembangan. Model Wilson, Oriondo, dan Antonio dimodifikasi dalam instrumen yang telah dibuat sebelumnya. Tahapan perancangan tes, uji coba tes, dan pengukuran tes merupakan tahapan yang terlibat dalam pembuatan instrumen dalam bentuk tes dengan memodifikasi Model Wilson dan Model Oriondo dan Antonio. (Oriondo, L.L.,& Dallo-Antonio, 1998). Berikut Matriks Instrumen HOTS Kimia yang dikembangkan.

Tabel 1. Matriks Instrumen HOTS Kimia

Materi Aspek/subaspek		Larutan Asam dan Basa		Stoikiometri Reaksi dan Titrasi Asam Basa			Larutan Penyangga			
		1.1	1.2	2.1	2.2	2.3	3.1	3.2	3.3	3.4
Menganalisis	A.1	1 dan 2		14			23			
	A.2		3 dan 4	15				24		
	A.3	6	5			16				25
Mengevaluasi	B.1		7 dan 8		17			26		
	B.2		9 dan 10			18			27	
Mencipta	C.1		11		19				28	
	C.2		12			20 dan 21			29	
	C.3	13				22				30

Berdasarkan Tabel 1, dapat dilihat bahwa butur yang dikembangkan berjumlah 30 butir soal dengan sebaran pada aspek menganalisis, mengevaluasi dan mencipta. Subjek penelitian adalah siswa kelas XI SMA Negeri di Yogyakarta. Jumlah sampel populasi diambil dengan tingkat kepercayaan 95% menggunakan acuan Tabel Morgan. (Krejcie, R. V., & Morgan, 1970). Berdasarkan tabel tersebut, 322 siswa merupakan jumlah sampel minimum yang perlu diambil jika jumlah populasi ± 2000 siswa. Sementara itu, 200 merupakan jumlah sampel minimum yang diperlukan untuk model 1-PL (model Rasch) (Wright, B. D., & Stone, 1979) atau 150-250 (Linacre, 1994). Faktor-faktor ini menyebabkan ditetapkannya 206 sebagai ukuran sampel penelitian.

Alat penilaian ini terdiri dari lima pilihan jawaban dengan justifikasi untuk setiap pertanyaan pilihan ganda. Dalam penelitian ini, koefisien validitas isi diperoleh dari penilaian penilaian ahli yang diberikan oleh ahli kimia dan pengukuran (ahli material) serta ahli konstruksi pengujian. Algoritma Aiken kemudian digunakan untuk memproses temuan penilaian ahli (Azwar, 1995) dengan rumus berikut:

$$V = \frac{\Sigma s}{n\,(c-1)} \quad (1)$$

$$:s = r - lo \quad (2)$$

lo = angka penilaian validitas yang terendah; c = angka penilaian validitas tertinggi ; dan r = angka yang diberikan oleh penilai. Koefisien validitas sekitar 0,7 masih dapat diterima dan dianggap memuaskan (Aiken, 1980).

Langkah pertama dalam analisis data adalah menggunakan perangkat lunak QUEST untuk mendeskripsikan fitur dan kelayakan instrumen tes kimia HOTS untuk kelas XI menggunakan teori respons butir. Analisis kecocokan atau analisis faktor eksploratori adalah dua metode untuk mengevaluasi asumsi unidimensional dalam analisis berdasarkan metodologi IRT. Sementara itu, kriteria yang berkisar dari -2 logit hingga +2 logit digunakan untuk menginterpretasikan fitur parameter butir dalam bentuk tingkat kesulitan butir. (Kriswantoro & Amelia, 2016) (Hambleton & Swaminathan, 1985). Sementara itu, karakteristik pada parameter butir berupa tingkat kesulitan butir diinterpretasi menggunakan kriteria dari (Baker, 2001) yakni

-2,00	-0,5	0	+0,5	+2
Sangat Mudah	Mudah	Sedang	Sukar	Sangat Sukar

HASIL DAN PEMBAHASAN

Para ahli mengevaluasi butir-butir soal yang dibuat, yang membuktikan validitas isi butir-butir soal. Rata-rata indeks Aiken sebesar 0,86 semakin mendukung kesimpulan bahwa semua butir soal dianggap valid berdasarkan hasil analisis, yang menunjukkan bahwa isi yang dinilai penting dan tercakup dalam kurikulum. Koefisien validitas, yang sekitar 0,86, masih dianggap memadai dan dapat diterima. (Aiken, 1980).

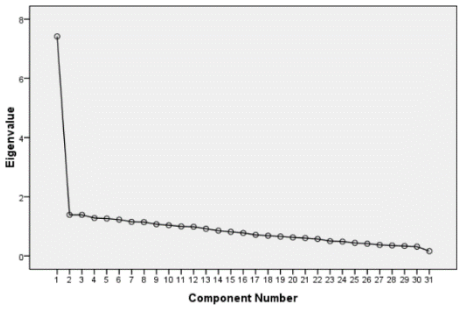
ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v14i4.2063

Dari 30 butir tes kimia HOTS memenuhi kriteria dalam analisis faktor dapat dilihat berdasarkan analisis yang telah dilakukan oleh peneliti bawasanya di dapat untuk nilai KMO sebesar 0,880 dan taraf signifikannya adalah 0,00. Dalam hal ini, kriteria untuk melanjutkan analisis dalam teori respon butir sudah bisa di lanjutkan. Lebih jelas dapat dilihat pada Tabel 2.

Tabel 2. *KMO and Bartlett's Test*

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.880
Bartlett's Test of Sphericity Approx. Chi Square	1703.772
Df	462
Sig.	.000

Selain dilihat dari nilai KMO nya, dilihat juga dari output lainnya seperti *Scree Plot*. Untuk membuktikan banyaknya dimensi yang terukur dalam suatu data, pada hasil scree plot dapat diamati banyaknya curam yang ada. Suatu item pada tes yang menilai lebih dari satu dimensi akan memerlukan kombinasi bakat yang berbeda dari peserta untuk menjawabnya. (Kriswantoro & Amelia, 2016). Jumlah dimensi ditunjukkan oleh kecuraman pergeseran, dan ketiadaan dimensi ditunjukkan oleh kerataan perubahan Nilai Eigen. (Retnawati, 2016). Berikut disajikan *screet plot unidimensi* pada Gambar 1.



Gambar 1. Scree Plot Uji Unidimensi

Uji unidimensi pada Gambar 1 dapat dijelaskan bahwa, nilai eigen yang pertama yaitu 7, kemudian turun drastis dari 7 menjadi 1, maka dapat dikatakan bahwa alat ukur tersebut mempunyai dimensi yang sangat kasat mata. maka dapat dikatakan bahwa alat ukur tersebut mempunyai dimensi yang sangat kasat mata. Dapat dikatakan bahwa instrumen ini mengukur dimensi tunggal atau perangkat bersifat unidimensi (Hambleton & Swaminathan, 1985). Tujuan dari persyaratan item satu dimensi adalah untuk mempertahankan invariansi teori respons item. (Kriswantoro & Amelia, 2016).

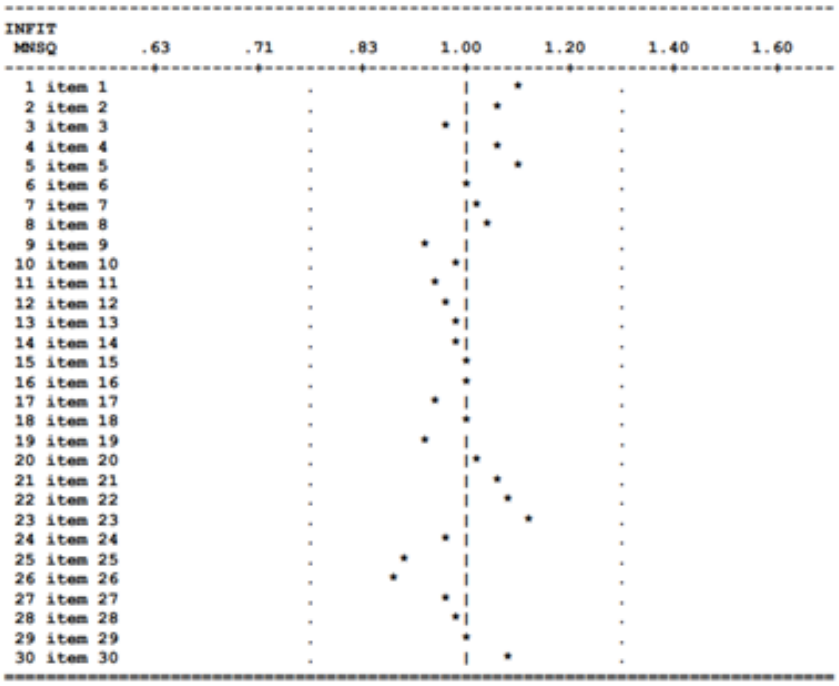
Memeriksa kecocokan setiap item terhadap model untuk melihat apakah cocok: jika suatu item cocok dengan model, nilai PT *Measure Corr*-nya positif, nilai *Outfit* MNSQ-nya antara 0,5 dan 1,5, dan nilai Outfit ZSTD-nya antara -2 dan +2 (Sumintono & Widhiarso, 2014). Berikut disajikan kecocokan butir pada Gambar 2.

Dari Gambar 2 dapat dijelaskan bahwa selama salah satu dari tiga kondisi terpenuhi, suatu objek dapat dianggap sesuai dengan model. Hal ini terlihat tidak hanya dari Outfit tetapi juga dari Infit MNSQ 0,77 hingga 1,3. Setiap satu dari tiga puluh butir pada instrumen ini sesuai dengan model. Perangkat lunak aplikasi winsteps digunakan untuk menghitung tingkat kesulitan. Soal yang paling sulit dan paling mudah dapat digunakan untuk menentukan tingkat kesulitan suatu item. Suatu butir dianggap sulit jika tingkat kesulitannya positif, dan mudah jika tingkat kesulitannya negatif. Tingkat kesulitan yang dihasilkan dimulai dari -0,88 dan naik hingga 1,04. Kecocokan yang lebih ketat antara -2 logit dan +2 logit menunjukkan item yang lebih mudah, sedangkan angka di antara keduanya menyiratkan item yang lebih sulit. Rumusan tingkat kesulitan peneliti tidak sesuai dengan tingkat kesulitan data empiris. Ini adalah hasil dari peneliti yang membuat item dan hanya menggunakan intuisinya untuk mengkategorikannya ke dalam tiga tingkat kesulitan: mudah, sedang, dan sulit. (Wang & Stanley, 1970). Karena mengukur tingkat kesulitan suatu soal ujian sebelum siswa mengerjakannya merupakan suatu tantangan, tidak selalu hal-hal yang dianggap "sulit" oleh peneliti dilihat dengan cara yang sama oleh siswa (Baker, 2001).

Karakteristik tingkat kesulitan setiap pertanyaan berkisar antara -0,88 dan 1,04. Pertanyaan nomor 6 memiliki tingkat kesulitan terendah (indeks: -0,88), sedangkan pertanyaan nomor 19 memiliki indeks kesulitan

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v14i4.2063

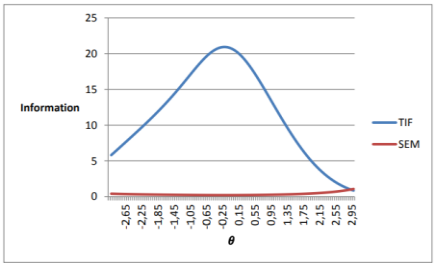
tertinggi (indeks: 1,04). Karakteristik pada parameter butir berupa tingkat kesulitan butir dari -2 sampai 2 dapat disimpulkan bahwa dari 30 butir soal tersebut memenuhi kriteria tingkat kesulitan yang baik. Hal ini disebabkan karena instrumen HOTS yang diujikan menuntut peserta didik untuk melakukan suatu proses berpikir dalam level kognitif yang lebih tinggi dengan mengintegrasikan berbagai mental kognitif berawal dari bernalar, kritis dalam mengolah informasi, menarik kesimpulan dan mengambil keputusan, serta kreatif untuk membuat berbagai strategi dalam melakukan pemecahan masalah (Herman et al., 2022).



Gambar 2. Map kecocokan *Item* Teradap Model

Tingkat kesukaran setiap butir bisa kita lihat dari hasil output, bawasannya pada setiap butir terdapat peluang menjawab benar dan menjawab salah pada setiap peserta didik. Peluang ini akan mempengaruhi hasil skor yang didapat dari hasil tes peserta didik. Pada aspek menganalisis, tingkat kesukaran paling tinggi terdapat pada butir no 14 yaitu sebesar 61,7%, dan tingkat kesukaran paling mudah terdapat pada soal no 3 yaitu 1,5%. Pada aspek mengevaluasi, tingkat kesukaran paling tinggi terdapat pada butir no 26 yaitu sebesar 70,4%, dan tingkat kesukaran paling mudah terdapat pada soal no 10 yaitu 2,9%. Pada aspek mencipta, tingkat kesukaran paling tinggi terdapat pada butir no 20 yaitu sebesar 57,3%, dan tingkat kesukaran paling mudah terdapat pada soal no 21 yaitu 2,9%. Suatu butir soal masuk dalam kategori sukar jika koefisiennya < 0,3; masuk dalam kategori sedang jika koefisiennya antara 0,3 sampai 0,7, dan masuk dalam kategori mudah jika koefisiennya > 0,7 (Kartowagiran et al., 2019).

Nilai fungsi informasi menunjukkan seberapa besar setiap butir soal berkontribusi pada penemuan bakat terpendam yang diukur oleh soal tersebut. SEM dan fungsi informasi memiliki hubungan terbalik. Kesalahan pengukuran berkurang dengan meningkatnya nilai fungsi informasi yang diberikan oleh butir soal. Jika butir soal memiliki fungsi informasi yang tinggi, nilai fungsi informasi perangkat soal akan tinggi. Hubungan antara Fungsi Informasi (IF) dan Pengukuran Kesalahan Baku (SEM) pada Gambar 3 ditunjukkan oleh kurva berikut.



Gambar 3. Fungsi Informasi dan *Standar Eror Meansurement* (SEM)

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v14i4.2063

Dari Gambar 3 dapat dijelaskan bahwa jika instrument HOTS Kimia diberikan kepada siswa dengan kemampuan menengah, ujian ini memiliki kesalahan pengukuran terkecil sebesar 0,21 dan akan menghasilkan informasi maksimum sebesar 20,92, atau -0,2. Skor kemampuan di mana grafik fungsi informasi dan grafik kesalahan pengukuran baku pada interval tersebut menunjukkan bahwa fungsi yang diperoleh tinggi atau kesalahan baku pengukuran informasi yang terjadi minimal mewakili batas bawah dan atas interval. Pemikiran akhir tentang fitur perangkat soal (cocok untuk siswa dengan kemampuan sedang).

SIMPULAN

Instrumen keterampilan berpikir tingkat tinggi untuk kimia memiliki kualitas yang menjadikannya instrumen yang memenuhi standar pengukuran. Berdasarkan data politomus empat kategori, instrumen ini telah memperoleh bukti empiris kesesuaian dengan *Partial Credit Model* (PCM) dan telah memenuhi persyaratan validitas konten dengan penilaian ahli. Instrumen keterampilan berpikir tingkat tinggi kimia memiliki semua itemnya dalam kriteria sangat baik. Instrumen keterampilan berpikir tingkat tinggi kimia sangat cocok untuk menilai kemampuan berpikir tingkat tinggi siswa dalam pelajaran kimia, karena didasarkan pada fungsi informasi.

Daftar Pustaka

Aiken, L. . (1980). *Content validity and realiability of single items or questionnaires*. Pepperdine University. https://doi.org/10.1177/001316448004000419

Amelia, R. N., & Kriswantoro, K. (2017). Implementation of Item Response Theory for Analysis of Test Items Quality and Students' Ability in Chemistry. *JKPK (Jurnal Kimia Dan Pendidikan Kimia)*, *2*(1), 1. https://doi.org/10.20961/jkpk.v2i1.8512

Azwar, S. (1995). Reliabilitas dan Validitas. *Buletin Psikologi*, *3*(1), 19–26.

Baker, F. . (2001). *The basics of item response theory (2nd Ed)*. ERIC Clearinghouse on Assessment and Evaluation. https://doi.org/10.1080/15366367.2018.1462078

Faradisa, B. T. Z. V., Kurniasih, S., & Berlian, L. (2024). Pengembangan Instrumen Tes 4TMC CBT Pada Materi Sistem Pernapasan untuk Mengukur Berpikir Kritis Siswa SMP Kelas VIII. *Jurnal Pendidikan MIPA*, *14*(4), 723–731. https://doi.org/https://doi.org/10.37630/jpm.v14i4.2010

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Kluwer-Nijjhoff Publish.

Herman, T., Hasanah, A., Nugraha, R. C., Harningsih, E., Ghassani, D. A., & Marasabessy, R. (2022). Pembelajaran Berbasis Masalah-High Order Thinking Skill (HOTS) pada Materi Translasi. *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *6*(1), 1131–1150. https://doi.org/10.31004/cendekia.v6i1.1276

Iskandar, D., & Senam, S. (2015). Studi Kemampuan Guru Kimia Sma Lulusan Uny Dalam Mengembangkan Soal Uas Berbasis Hots. *Jurnal Inovasi Pendidikan IPA*, *1*(1), 65. https://doi.org/10.21831/jipi.v1i1.4533

Isnawati, I., Anfa, Q., & Rohmani, L. A. (2024). Analisis Keterampilan Berpikir Kritis Siswa Kelas IX SMP dalam Menyelesaikan Soal Berbasis Higher Order Thinking Skills pada Materi Sistem Reproduksi Manusia. *Jurnal Pendidikan MIPA*, *14*(3), 723–731. https://doi.org/10.37630/jpm.v14i3.1754

Istiyono, E., Mardapi, D., & Suparno, S. (2014). PENGEMBANGAN TES KEMAMPUAN BERPIKIR TINGKAT TINGGI FISIKA (PysTHOTS) PESERTA DIDIK SMA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *18*(1), 1–12. https://doi.org/10.21831/pep.v18i1.2120

Johar, R. (2012). Domain Soal PISA untuk Literasi Matematika. *Jurnal Peluang*, *1*(1), 30.

Kartowagiran, B., Mardapi, D., Purnama, D. N., & Kriswantoro, K. (2019). Parallel tests viewed from the arrangement of item numbers and alternative answers. *REID (Research and Evaluation in Education)*, *5*(2), 169–182. https://doi.org/10.21831/reid.v5i2.23721

Krejcie, R. V., & Morgan, D. W. (1970). Determining Sample Size For Research Activities. *Educational and Psychological Measurement, 30 (3)*, 607–610. https://doi.org/10.1177/001316447003000308

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v14i4.2063

Kriswantoro;, Kartowagiran, B., & Rohaeti, E. (2021). A Critical Thinking Assessment Model Integrated with Science Process Skills on Chemistry for Senior High School. *European Journal of Educational Research*, *10*(1), 285–298. https://doi.org/10.12973/eu-jer.10.1.285

Kriswantoro, & Amelia, R. N. (2016). Peningkatan Kompetensi Calon Pendidik Kimia Melalui Item Response Theory: Strategi Menghadapi Masyarakat Ekonomi Asean. *SEMINAR NASIONAL KIMIA DAN PENDIDIKAN KIMIA VIII*, 64–73.

Linacre, J. M. (1994). Sample Size and Item Calibration or Person Measure Stability. *Rasch Measurement Transactions*, *7*(4), 328. http://www.rasch.org/rmt/rmt74m.htm

Oriondo, L.L.,& Dallo-Antonio, E. M. (1998). *Evaluation Educational Outcomes*. Rex Printing Compagny, inc.

Pratiwi, L. F. (2022). Analisis Kemampuan Guru Dalam Membuat Soal Tipe Hots (High Order Thinking Skills) Mata Pelajaran Matematika. *Humantech Jurnal Ilmiah Multi Disiplin Indonesia*, *1*(6), 765–771.

Putri, V. A., & Syolendra, D. F. (2024). Meta-Analisis : Pengaruh Model Problem Based Learning Terhadap Hasil Belajar Kimia Peserta Didik Meta-Analysis : The Problem- Based Learning Model ' s Effect on Students '. *Edukimia*, *6*(2), 75–80. https://doi.org/10.24036/ekj.v6.i2.a

Resnick, L. B. (1987). Education and Learning to Think. In *Education and Learning to Think* (Issue January 1987). https://doi.org/10.17226/1032

Retnawati, H. (2016). *Validitas reliabilitas & karakteristik butir (panduan untuk peneliti,mahasiswa, dan psikometri)*. Prama Publising.

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi Model Rasch Untuk Penelitian Ilmu-Ilmu Sosial*. trim Komunikata Publishing House.

Sunggingwati, D., & Nguyen, H. T. M. (2013). Teachers' questioning in reading lessons: A case study in Indonesia. *Electronic Journal of Foreign Language Teaching*, *10*(1), 80–95.

Sunyono;, Wirya, I. W., Suyanto, E., & Suyadi, G. (2009). Identifikasi masalah kesulitan dalam pembelajaran kimia SMA kelas X di propinsi Lampung. *Journal Pendidikan MIPA (JPMIPA)*, *10*(2), 9–18.

Wang, M. W., & Stanley, J. C. (1970). Differential weighting: a review of methods and empirical studies. *Review of Educational Research.*, *40*(5), 663–705.

Wright, B. D., & Stone, M. H. (1979). Rasch Rating Scale Analysis of the Arabic Version of the Physical Activity Self-Efficacy Scale for Adolescents: A Social Cognitive Perspective. *Psychology*, *6*(16).