Jurnal Ilmiah Kanderang Tingang
Vol.14 No.1 Januari-Juni 2023
FKIP Universitas Palangka Raya

ISSN 2087-166X (printed)
ISSN 2721-012X (online)
DOI: https://doi.org/10.37304/jikt.v14i1.180

Aplikasi Andrich Rating Scale Model Pada Analisis Psikometrik Tes Uraian Kimia Dasar I

Rizki Nor Amelia[(1)], Anggi Ristiyana Puspita Sari[(2)], Sri Rejeki Dwi Astuti[(3)], Dian Normalitasari Purnama[(4)]

[1]Prodi Pendidikan IPA, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Semarang, Indonesia

[2]Prodi Pendidikan Kimia, Fakultas Ilmu Pendidikan, Universitas Palangka Raya, Indonesia

[3]Prodi Pendidikan Kimia, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Yogyakarta, Indonesia

[4]Prodi Pendidikan Akuntansi, Fakultas Ekonomi, Universitas Negeri Yogyakarta, Indonesia

Email Author: rizkinoramelia@mail.unnes.ac.id

Diterima:07-12-2022; Diperbaiki:08-01-2023; Disetujui:10-01-2023

ABSTRAK

Kimia Dasar I merupakan mata kuliah wajib yang ditempuh oleh calon Guru IPA dimana penguasaan terhadap mata kuliah tersebut dapat digali melalui tes uraian. Tujuan penelitian ini adalah mendeskripsikan karakteristik psikometrik tes uraian Kimia Dasar I menggunakan Andrich Rating Scale Model dengan bantuan program Winsteps. Penelitian yang dilakukan pada Semester Gasal Tahun Akademik 2022/2023 melibatkan 46 mahasiswa yang terpilih melalui teknik cluster random sampling. Meskipun hasil analisis psikometrik menunjukkan bahwa rating scale belum berfungsi sebagai mana mestinya, namun instrumen tes uraian Kimia Dasar I terbukti memiliki validitas konstruk yang baik (unidimensi) dengan keseluruhan butir berada pada kategori tingkat kesukaran sedang. Penelitian ini sekaligus membuktikan bahwa tidaklah mudah membuat dan menetapkan kategori dalam tes uraian. Penggunaan Andrich Rating Scale Model memberikan informasi yang sangat bermanfaat untuk menggambarkan dan memperbaiki kualitas psikometrik tes uraian pada pengukuran kemampuan Kimia Dasar I calon guru IPA.

Kata kunci: *andrich rating scale model, tes uraian, analisis psikometrik, kimia dasar I*

PENDAHULUAN

Pasal 10 ayat (1) Undang-Undang Nomor 14 Tahun 2005 tentang Guru dan Dosen mengamanatkan bahwa Guru harus memiliki kompetensi pedagogik, kompetensi kepribadian, kompetensi sosial, dan kompetensi profesional, yang keempatnya bersifat holistik dan merupakan suatu kesatuan dari ciri guru yang profesional. Guna mewujudkan dan memenuhi kebutuhan pendidik yang profesional dan berkompeten tersebut, dibentuklah suatu lembaga bernama Lembaga Pendidikan Tenaga Kependidikan (LPTK), dimana lembaga ini terdiri atas perguruan tinggi dibawah naungan Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi serta Kementerian Agama, yang diberi tugas oleh pemerintah untuk menyelenggarakan program pengadaan guru pada pendidikan anak usia dini jalur pendidikan formal, pendidikan dasar, dan/atau pendidikan menengah serta untuk menyelenggarakan dan mengembangkan ilmu kependidikan dan nonkependidikan (Pasal 1 ayat 3, Permenristekdikti No 55 Tahun 2017 tentang Standar Pendidikan Guru). Sebagai lembaga pencetak guru, LPTK berpedoman pada Permenristekdikti No 55 Tahun 2017 tentang Standar Pendidikan Guru.

ISSN 2087-166X (printed)
ISSN 2721-012X (online)
DOI: https://doi.org/10.37304/jikt.v14i1.180

Dewasa ini, Indonesia dihadapkan pada empat tantangan besar, yakni bonus demografi, globalisasi, pluralisme, dan revolusi industri 4.0 yang menuntut LPTK sebagai institusi yang menghasilkan pendidik generasi bangsa memiliki tanggung jawab yang sangat tinggi (Arifian, 2019). Salah satu langkah yang dapat ditempuh untuk menjawab tantangan tersebut adalah melakukan pembenahan intrakurikuler. Pembenahan ini dapat dilakukan melalui dua cara. Pertama, penataan dengan cermat standar isi kurikulum dan standar evaluasi untuk melihat ketercapaian tujuan pembelajaran yang dihubungkan dengan standar isi dan proses, dan kedua, penyelarasan kurikulum dengan Kerangka Kualifikasi Nasional Indonesia (KKNI) sesuai Peraturan Presiden Republik Indonesia Nomor 8 Tahun 2012 (Arifian, 2019). Kurikulum merupakan "jantung" institusi pendidikan atau sistem pembelajaran. Tanpa kurikulum proses pembelajaran menjadi tidak jelas arah dan orientasinya (Maslhahah, 2018). Dengan penyelarasan kurikulum dalam konteks KKNI, artinya LPTK telah menyesuaikan pendidikan sebagai pemasok Sumber Daya Manusia dengan dunia kerja yang memiliki kebutuhan dan tuntutan yang dinamis (Maslhahah, 2018) serta memiliki acuan kesetaraan kompetensi di tengah kebervariasaian persoalan mutu pendidikan di dalam negeri dan di tengah tuntutan dan persaingan mutu sumber daya manusia tingkat global (Payong, 2015).

Kimia dasar adalah salah satu mata kuliah wajib dalam struktur kurikulum Universitas Negeri Semarang yang harus ditempuh calon guru Ilmu Pengetahuan Alam (IPA). Mata kuliah ini bertujuan untuk membekali penguasaan teori dan praktik mengenai Materi, Karakteristik, dan Pengukurannya; Atom, Molekul, dan Ion; Teori Atom dan SPU, Konsep Mol dalam Zat dan Larutan; Stoikiometri dalam Reaksi Kimia; Termokimia; Ikatan Kimia dan Bentuk Geometri Molekul; Tinjauan Fase Gas, Cair dan Padat; Kimia Unsur, Kimia Inti; Kimia Organik dan Biokimia. Salah satu cara yang dapat digunakan untuk mengukur penguasaan terhadap mata kuliah ini adalah melalui tes uraian. Tes uraian merupakan salah satu teknik penilaian tertua, yang telah ada jauh sebelum tes pilihan ganda, dan merupakan bagian utama dari sebuah pendidikan formal (Allanson & Notar, 2019). Pada dasarnya tidak ada alat penilaian yang sempurna, tidak terkecuali tes uraian, namun jenis tes ini menawarkan berbagai keuntungan kepada pembuat tes ketika bentuk pilihan ganda, benar-salah, atau format serupa lainnya tidak cukup untuk menggali informasi yang dibutuhkan. (Boye, 2019).

Tes uraian dibagi menjadi dua jenis, yakni *extended response* (*open ended*) dan *restricted response* (Allanson & Notar, 2019). Uraian *extended response* biasanya digunakan untuk mengukur hasil belajar yang meliputi mengungkapkan, mengorganisasikan, menganalisis, mengevaluasi ide, menghubungkan, dan mengintegrasikan ide dengan informasi faktual lainnya (Linn & Miller, 2005), sedangkan uraian *restricted response* biasanya digunakan untuk menilai tujuan pembelajaran taksonomi Bloom yang rendah seperti membuat daftar, mendefinisikan, atau sekedar menggambarkan (Reynolds, Livingston, & Wilson, 2006). Ditinjau dari kemudahan penilaian, *restricted response* memang memungkinkan pendidik menilai hasilnya dengan mudah karena adanya pembatasan terhadap tanggapan, sehingga skor tanggapan juga cenderung lebih reliabel (Linn & Miller, 2005).

Umumnya tes uraian dipilih karena tes ini menantang siswa untuk memberikan respons dibandingkan sekedar memilih respons (Reiner, Bothell,

ISSN 2087-166X (printed)
ISSN 2721-012X (online)
DOI: https://doi.org/10.37304/jikt.v14i1.180

Sudweeks, & Wood, 2002). Meskipun tetap tidak kebal terhadap perilaku tebak-tebakan (Clay, 2001), tes uraian lebih mampu mengungkap kemampuan berpikir tingkat tinggi (Nilson, 2017) maupun hasil belajar pada materi yang lebih kompleks (Minbashian, Huon, & Bird, 2004), memfasilitasi penilaian yang lebih otentik (Wiggins, 2011), melatih kemampuan komunikasi (proses berpikir, mengorganisasi dan logika) hingga membuat siswa membangun strategi belajar lebih mendalam (Nilson, 2017). Dari segi waktu konstruksi, tes uraian dapat dibuat lebih cepat dibandingkan tes pilihan ganda (Clay, 2001; Nilson, 2017). Namun itu tidak berarti pertanyaan uraian yang bagus mudah dikonstruksi. Makna "lebih mudah" ini dalam arti relatif, artinya konstruksi pertanyaan uraian tetap membutuhkan waktu dan usaha (Reiner, Bothell, Sudweeks, & Wood, 2002). Salend (2011) menegaskan bahwa tes yang dirancang dengan tergesa-gesa dapat berdampak negatif pada pertumbuhan akademik.

Tantangan lain dalam konstruksi tes uraian adalah pemilihan skala penilaian (*rating scale*) yang tepat dan penetapan kriteria berdasarkan tujuan evaluasi (Bacha, 2001) yang umum terangkum dalam pola penskoran. Pola penskoran merupakan salah satu faktor yang diklaim bertanggungjawab pada ketidakandalan (Ebuoh, 2018) karena kesesuaian antara jenis penskoran dengan indikator yang diukur berimplikasi pada validitas dan reliabilitas informasi yang menggambarkan hasil belajar siswa (Wahyuni, Gumela, & Maulana, 2020). Setidaknya terdapat dua jenis pola penskoran yang dapat dipilih, yakni penskoran analitik (berupa daftar elemen utama yang diharapkan ada pada respon siswa, lebih cocok diaplikasikan pada *restricted response*) (Ebuoh, 2018) dan penskoran holistik (mempertimbangkan seluruh tanggapan tertulis, lebih ringkas karena tidak memasukkan kriteria evaluasi secara terperinci, dan lebih cocok diaplikasikan pada *extended response*) (Ghalib & Al-Hattami, 2015).

Terkait *rating scale*, Andrich Rating Scale Model (Andrich RSM) merupakan salah satu model dalam Teori Tes Modern yang dapat dimanfaatkan untuk menentukan keberfungsian *rating scale*, baik dari segi jumlah opsi dan labelnya (Zile-Tamsen, 2017). Konsep Andrich RSM bekerja untuk data ordinal pada *rating scale* dengan memperkirakan nilai abilitas responden dan lokasi (tingkat kesukaran) butir pada suatu skala pengukuran interval yang dikenal sebagai logits (*logarithm of odds*) (Chong, Mokshein, & Mustapha, 2022). Di Indonesia, penelitian yang menggali karakteristik psikometrik tes uraian Kimia Dasar I menggunakan Andrich RSM cukup sulit ditemukan, umumnya karakteristik hanya dieksplorasi menggunakan Teori Tes Klasik. Padahal sudah jelas Teori Tes Klasik memiliki berbagai kelemahan dan sudah mulai ditinggalkan (Auné, Abal, & Attorresi, 2020). Untuk itu, penelitian ini dilakukan untuk mendeskripsikan psikometrik dari instrumen tes uraian Kimia Dasar I menggunakan Andrich Rating Scale Model.

METODOLOGI PENELITIAN

Secara umum, penelitian ini merupakan penelitian deskriptif kuantitatif. Penelitian ini dilakukan pada Semester Gasal Tahun Akademik 2022/2023. Populasi dalam penelitian ini adalah 92 mahasiswa Semester I (Angkatan 2022) Program Studi Pendidikan IPA Universitas Negeri Semarang yang mendapatkan mata kuliah Kimia Dasar I dan terbagi ke dalam 4 rombongan belajar. Dari

ISSN 2087-166X (printed)
ISSN 2721-012X (online)
DOI: https://doi.org/10.37304/jikt.v14i1.180

populasi ini, digunakan teknik *cluster random sampling*, sehingga terpilih sampel sebanyak dua rombel (N = 46, Perempuan = 35; Laki-laki = 11 mahasiswa).

Instrumen penelitian yang digunakan berupa tes uraian yang terdiri atas 10 butir pertanyaan (skor total minimum = 0 dan skor total maksimum = 9) dengan spesifikasi sebagaimana dalam Tabel 1.

Tabel 1. Spesifikasi Instrumen Tes Uraian Kimia Dasar I

No	Materi	Indikator	Dimensi Kognitif	Tingkat Kesukaran	Butir soal ke-
1.	Materi, Karakteristik, dan Pengukurannya	Diberikan berbagai data sifat senyawa dan campuran, mahasiswa dapat mengklasifikasikan dasar perbedaan sifat yang dimiliki oleh senyawa dan campuran tersebut	C3	sedang	1
		Diberikan data komponen senyawa pada gula hasil pengolahan tebu, mahasiswa dapat memutuskan teknik pemisahan campuran yang tepat disertai alasannya	C5	sukar	2
2.	Teori Atom dan Sistem Periodik Unsur	Diberikan grafik sifat keperiodikan unsur dalam satu golongan dan satu periode, mahasiswa dapat menganalisis kecenderungan afinitas elektron atom-atom dalam satu golongan	C4	sukar	3
		Diberikan data nomor atom beberapa unsur yang terdapat dalam cangkang telur, mahasiswa dapat menentukan letaknya dalam Sistem Periodik Unsur	C3	sedang	4
		Diberikan data energi ionisasi unsur-unsur periode tiga yang secara umum naik, mahasiswa dapat menganalisis mengapa ada penyimpangan pada unsur alumunium dan silikon	C4	sukar	6
3.	Ikatan kimia dan bentuk geometri molekul	Diberikan data nomor atom beberapa unsur dalam suatu molekul, mahasiswa dapat mengkorelasikan tipe bentuk molekul terhadap bentuk geometri molekulnya berdasarkan teori VSEPR	C4	sukar	5
4.	Konsep Mol dalam Zat dan Larutan	Diberikan data-data untuk membuat suatu larutan basa kuat, mahasiswa dapat menghitung konsentrasi larutan yang dibuat tersebut	C3	sedang	7
		Diberikan data mol suatu senyawa, mahasiswa dapat menghitung jumlah partikel yang terkandung dalam senyawa tersebut	C3	sedang	8
5.	Stoikiometri dalam reaksi kimia	Diberikan data untuk melakukan reaksi pembakaran glukosa menggunakan oksigen, mahasiswa dapat mengitung massa oksigen yang bereaksi	C3	sedang	9
		Diberikan data yang dibutuhkan untuk mengetahui konsentrasi suatu asam kuat pekat, mahasiswa dapat menghitung volume yang harus diambil untuk membuat larutan yang	C3	sedang	10

ISSN 2087-166X (printed)
ISSN 2721-012X (online)
DOI: https://doi.org/10.37304/jikt.v14i1.180

No	Materi	Indikator	Dimensi Kognitif	Tingkat Kesukaran	Butir soal ke-
		diencerkan dari asam kuat pekat tersebut			

Data berupa respon jawaban politomus diolah menggunakan program Winsteps. Formulasi Andrich untuk Rating Scale Model disajikan dalam persamaan (1). Pada persamaan tersebut, θ_n and δ_i berturut-turut adalah parameter abilitas responden ke-n dan parameter lokasi (tingkat kesukaran) butir ke-i, τ_j adalah parameter lokasi dari langkah *j* pada setiap butir dan *k* menunjukkan kategori (Andrich, 1978)

$$P_{xni} = \frac{exp\left[-\sum_{j=0}^{x}\tau_j + x(\theta_n - \delta_i)\right]}{\sum_{k=0}^{m} exp\left[-\sum_{j=0}^{x}\tau_j + x(\theta_n - \delta_i)\right]} \ldots\ldots (1)$$

HASIL PENELITIAN DAN PEMBAHASAN

Deskripsi karakteristik psikometrik tes uraian Kimia Dasar I dimulai dari pembuktian validitas konstruk (unidimensi). Pada Teori Tes Modern, dimensi adalah asumsi kunci yang menunjukkan jumlah sifat laten yang menentukan respon butir (Chou & Wang, 2010). Asumsi ini sangat penting untuk memastikan bahwa semua butir dalam tes mengukur sifat laten yang sama (Apple, 2013). Analisis unidimensi dengan Winsteps menggunakan default *Principal Component Analysis* (PCA). Dalam ukuran unidimensi, varians yang diamati diharapkan dapat dijelaskan dengan langkah-langkah yang cocok dengan varians yang diharapkan oleh model (Ee, Yeo, & Mohd Kosnin, 2018). Umumnya digunakan kriteria minimum pada *raw variance explained by measures* sebesar 40% dan *unexplained variance in the first factor* seharusnya tidak boleh lebih dari 15% (Conrad, Conrad, Passetti, Funk, & Dennis, 2015; Fisher, 2007). Berdasarkan kriteria tersebut, Tabel 2 menunjukkan bahwa asumsi unidimensi telah terpenuhi (*raw variance explained by measures* = 62.7% dan *unexplained variance in the first factor* = 7.9%) atau validitas konstruk telah terbukti.

Tabel 2. Varians Residu Standar (Eigenvalue unit) pada instrumen Kimia Dasar I

		Empirical		Modeled
Total raw variance in observations	= 44.6	100.0%		100.0%
Raw variance explained by measures	= 29.6	66.4%		66.7%
Raw variance explained by persons	= 18.0	40.3%		40.5%
Raw variance explained by items	= 11.6	26.0%		26.2%
Raw unexplained variance (total)	= 15.0	33.6%	100.0%	
Unexplained variance in 1st contrast	= 6.1	13.6%	40.5%	
Unexplained variance in 2nd contrast	= 2.3	5.1%	15.0%	
Unexplained variance in 3rd contrast	= 1.7	3.9%	11.5%	
Unexplained variance in 4th contrast	= 1.6	3.5%	10.4%	
Unexplained variance in 5th contrast	= 1.2	2.7%	8.1%	

Karakteristik psikometrik selanjutnya yang dieksplorasi adalah reliabilitas. Hasil analisis menunjukkan jika instrumen tes uraian Kimia Dasar I telah memiliki keandalan yang baik, terbukti dari koefisien *model person reliability* dan *real person reliability* berturut-turut sebesar 0.80 dan 0.62; serta koefisien *item model reliability* dan *item real reliability* keduanya sebesar 0.96. Sementara itu,

ISSN 2087-166X (printed)
ISSN 2721-012X (online)
DOI: https://doi.org/10.37304/jikt.v14i1.180

kecocokan butir (*item fit*) digunakan untuk mengevaluasi apakah masing-masing butir berkontribusi pada unidimensionalitas (Conrad, et.al., 2015) dan sejauhmana pola sampel respon terhadap suatu butir itu konsisten seperti respon orang lain dalam menanggapi butir-butir yang lain (Razak, Khairani, & Thien, 2012). Setelah diidentifikasi, butir yang terdeteksi misfit dapat diperiksa secara kualitatif untuk menentukan penyebab masalah, apakah susunan kata yang membingungkan atau butir tersebut justru mengukur konstruk yang berbeda dari konstruk utama yang diukur (Conrad, et.al., 2015). Adapun kriteria yang dapat diacu adalah $0.5 \leq$ outfit MNSQ (*outlier-sensitive or information-weighted fit Mean Square*) ≤ 1.5 (Linacre, 2021).

Tingkat kesukaran butir (*b*) sering disebut sebagai parameter lokasi karena menunjukkan titik dalam skala kemampuan yang memiliki peluang 50% untuk dijawab dengan benar (Mahmud, 2017) dan biasanya berkisar dari -2.5 hingga +2.5 logit (Meijer & Tendeiro, 2018). Dari range tersebut, dapat dikategorikan sebagai berikut: range butir mudah $-2.5 \leq b < -1.0$; range butir sedang $-1.0 \leq b < +1.0$; dan range butir sukar $1.0 \leq b \leq +2.5$. Dalam kurva karakteristik butir, butir sukar berada di kanan skala yang mengindikasikan semakin tinggi kemampuan yang dibutuhkan untuk merespon dengan benar, sedangkan butir yang lebih mudah berada di sebelah kiri skala. Hasil analisis item fit dan tingkat kesukaran butir disajikan pada Tabel 3. Berdasarkan tabel tersebut, terindikasi 40% item misfit dan keseluruhan butir berada dalam kategori tingkat kesukaran sedang.

Tabel 3. Karakteristik Psikometrik Butir Uraian Kimia Dasar I

Butir No	OUTFIT MNSQ	b
1	2.50	-0.75
2	0.44	0.58
3	1.04	-0.11
4	0.79	0.21
5	0.67	0.65
6	2.67	-0.48
7	1.48	-0.41
8	1.01	-0.04
9	1.10	0.74
10	1.66	-0.39

Karakteristik psikometrik terakhir yakni kategorisasi Andrich RSM yang disajikan pada Tabel 4. Terdapat lima kriteria yang dapat digunakan untuk memaknai hasil analisis dari Tabel 4. Pertama, setiap kategori memiliki minimal 10 observasi. Pada aspek ini, tidak semuanya terpenuhi karena hanya kategori 3, 6, dan 9 yang memiliki *observed count* lebih dari 10. Oleh sebab itu, dapat disimpulkan jika secara umum, frekuensi observasi tidak cukup digunakan untuk melakukan estimasi *rating scale* yang stabil. Kedua, *observed average* mengalami peningkatan secara monoton seiring dengan meningkatnya kategori. Pada kategori 1 didapatkan *observed average* sebesar -1.28 logit, ini berarti rata-rata perkiraan abilitas untuk semua responden yang menjawab kategori 1 pada butir apapun dalam instrumen tes uraian Kimia Dasar I juga -1.28 logit. Nilai ini harus meningkat seiring dengan meningkatnya kategori respons karena menunjukkan bahwa responden yang memiliki abilitas tinggi akan mendukung kategori yang lebih tinggi secara progresif, dan demikian pula sebaliknya (Bond & Fox, 2015).

ISSN 2087-166X (printed)
ISSN 2721-012X (online)
DOI: https://doi.org/10.37304/jikt.v14i1.180

Pada aspek ini, secara umum terjadi peningkatan, kecuali pada kategori 2 dan kategori 8 yang justru mengalami penurunan.

Ketiga, nilai OUTFIT MNSQ dibawah 2.0. Untuk kriteria ini, outfit MNSQ berkisar pada 0.02-3.30 dimana kategori 4 dan 6 tidak memenuhi kriteria yang dipersyaratkan. Tingginya nilai OUTFIT pada kedua kategori tersebut mengindikasikan bahwa kedua kategori memberikan lebih banyak *noise* dibandingkan makna pengukuran. Kategori semacam ini memerlukan penyelidikan empiris lebih lanjut apakah lebih baik disederhanakan atau tidak (Bond & Fox, 2015). Hal ini selaras dengan Andrich & Luo (2002) yang menyebutkan jika terlalu banyak kategori atau pendefinisian kategori yang kurang sesuai dapat menjadi sumber utama yang menyebabkan *disordered category* yang pada akhirnya berimplikasi pada item misfit.

Keempat, *Step Callibration* pada *Andrich Treshold* mengalami peningkatan secara monoton seiring dengan meningkatnya kategori. Secara umum hasil analisis pada kriteria ini juga menunjukkan jika tidak ada konsistensi peningkatan *Andrich Treshold* seiring dengan meningkatkanya kategori, misalnya kategori 2 yang terdeteksi lebih mudah daripada kategori 1 atau kategori 7 yang terdeteksi lebih sukar daripada kategori 9. Sama halnya dengan kriteria ketiga, ketika ambang batas kategori mengalami gangguan, masalahnya sering kali karena memiliki terlalu banyak opsi respons, dan ini biasanya dapat diselesaikan dengan menciutkan respons dengan pertimbangan tertentu (Andrich & Luo, 2002). Terakhir, Selisih (*width*) *Step Callibration* pada *Andrich Treshold* berada pada range 1.4-5.0 logit. Konsisten dengan sebelumnya, secara umum untuk kriteria ini juga kurang terpenuhi.

Tabel 4. Struktur Kategorisasi *Andrich Rating Scale Model*

Category	Observed Count (%)	Observed Average	OUTFIT MNSQ	Andrich Treshold	Thresholds between category (width)
1	4 (1)	-1.28	0.97	NONE	
2	2 (0)	-1.54*	0.36	0.29	Category 1-2 (0.29)
3	223 (48)	0.17	1.27	-4.71	Category 2-3 (5.00)
4	6 (1)	0.48	3.30	3.87	Category 3-4 (8.55)
5	7 (2)	0.54	0.39	0.31	Category 4-5 (-3.58)
6	36 (8)	0.80	2.30	-1.00	Category 5-6 (-1.31)
7	3 (1)	0.97	0.02	3.28	Category 6-7 (4.28)
8	8 (2)	0.89*	0.34	-0.04	Category 7-8 (-3.32)
9	171 (37)	1.12	1.08	-1.99	Category 8-9 (-1.95)

Hasil analisis pada lima kriteria keberfungsian *rating scale* menunjukkan bahwa bukti empiris pada kategori *rating scale* instrumen tes uraian Kimia Dasar I belum dapat berfungsi sebagaimana mestinya. Oleh sebab itu, solusi yang dapat diambil adalah melakukan penyederhanaan kategori (terutama pada kategori yang berdekatan) untuk meningkatkan arti dan fungsi *rating scale* dalam suatu pengukuran (Linacre, 2002).

KESIMPULAN

Ditinjau dari validitas konstruk (unidimensionalitas), instrumen tes uraian Kimia Dasar I sudah terbukti memiliki validitas konstruk yang baik. Demikian pula jika ditinjau dari aspek reliabilitas person maupun reliabilitas item nya. Sementara itu, masih ada beberapa butir yang tidak cocok dengan Andrich RSM,

ISSN 2087-166X (printed)
ISSN 2721-012X (online)
DOI: https://doi.org/10.37304/jikt.v14i1.180

artinya butir-butir ini perlu direvisi kembali dibandingkan harus dibuang. Analisis tingkat kesukaran butir secara empiris memberikan hasil 100% butir berada pada kategori tingkat kesukaran sedang. Hal ini sedikit berbeda dengan tingkat kesukaran teoritis yang di judgement bahwa tingkat kesukaran butir menyebar dari mudah, sedang, hingga sukar.

Pada akhirnya, analisis Andrich RSM menyimpulkan bahwa *rating scale* yang digunakan dalam instrumen tes kimia dasar belum dapat berfungsi dengan baik. Judgement 9 kategori yang diberikan dalam menilai langkah-langkah pengerjaan soal Kimia Dasar I ternyata tidak konsisten, sehingga mengakibatkan kerancuan dalam interpretasi abilitas. Penelitian ini juga membuktikan bahwa tidaklah mudah membuat dan menetapkan kategori dalam tes uraian. Penggunaan Andrich RSM memberikan informasi yang sangat bermanfaat dalam melakukan analisis psikometrik, sehingga kedepannya dapat dilakukan perbaikan agar kemampuan kimia dasar calon guru IPA dapat tergambarkan dengan tepat.

DAFTAR RUJUKAN

Allanson, P.E., & Notar, C.E. 2019. Design, construction, grading of essay questions 1.0 for teachers. *American International Journal of Humanities and Social Science, 5*(3), 1-11.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Andrich, D. & Luo, G. 2002. Conditional pairwise estimation in the Rasch model for ordered response categories using principal component. *J Appl Meas, 4*, 205-221.

Apple, M.T. 2013. Using rasch analysis to create and evaluate a measurement instrument for foreign language classroom speaking anxiety. *JALT Journal, 35*(1), 5-28.

Arifian, F.D. 2019. Peran lembaga pencetak tenaga kependidikan (LPTK) dalam mempersiapkan generasi emas bangsa. *Jurnal Pendidikan dan Kebudayaan Missio, 11*(1), 26-38.

Auné, S. E., Abal, F. J. P., & Attorresi, H. F. 2020. Análisis psicométrico mediante la Teoría de la Respuesta al Ítem: modelización paso a paso de una Escala de Soledad. *Ciencias Psicológicas, 14*(1), 1–15. https://doi.org/10.22235/cp.v14i1.2179

Bacha, N. 2001. Writing evaluation: What can analytic versus holistic essay scoring tell us?. *System, 29*, 371-383

Bond, T.G., & Fox, C.M. 2015. *Applying the rasch model: Fundamental mesurement in the human sciences (Third ed).* New York: Routledge.

Boye, A.P. 2019. *Writing better essay exams*. Manhattan: IDEA Paper.

Chong, J., Mokshein, S.E., & Mustapha, R. 2021. Applying the rasch rating scale model (RSM) to investigate the rating scales function in survey research instrument. *Cakrawala Pendidikan, 41*(1), 97-111. https://doi.org/10.21831/cp.v41i1.39130

Chou, Y.T., & Wang, W.C. 2010. Checking dimensionality in item response models with principal component analysis on standardized residuals. *Education and Psychological Measurement, 70*(5), 717-731.

Clay, B. 2001. *Is this a trick question? A short guide to writing effective test questions*. Lawrence, KS: Kansas Curriculum Center.

ISSN 2087-166X (printed)
ISSN 2721-012X (online)
DOI: https://doi.org/10.37304/jikt.v14i1.180

Conrad, K.M., Conrad, K.J., Passetti, L.L., Funk, R.R., & Dennis, M.L. 2015. Validation of the full and short-form self-help involvement scale against the rasch measurement model. *Eval Rev*, *39*(4), 395-427. https://doi.org/10.1177/0193841X15599645.

Ebuoh, C.N. 2018. Effects of analytical and holistic scoring patterns on scorer reliability in biology essay tests. *World Journal of Education, 8*(1), 111-117. https://doi.org/10.5430/wje.v8n1p111

Ee, Ng.Sar., Yeo, K.J., & Mohd Kosnin, A.bt. 2018. Item analysis for the adopted motivation scale using rasch model. *International Journal of Evaluation and Research in Education*, *7*(4), 264-269. https://doi.org/10.11591/ijere.v7.i4.pp264-269

Fisher, W. 2007. Rating scale instrument quality criteria. *Rasch Measurement Transaction, 21*, 1095-1098.

Ghalib, T.K., & Al-Hattami, A.A. 2015. Holistic versus analytic evaluation of EFL writing: A case study. *English Language Teaching, 8*(7), 225-236.

Indonesia. *Undang-Undang Nomor 14 Tahun 2005 tentang Guru dan Dosen*. Lembaran Negara RI Tahun 2005 Nomor 157, Tambahan Lembaran Negara Republik Indonesia Nomor 4586. Sekretariat Negara. Jakarta.

Linacre. J.M. 2021. *A user's guide to WINSTEPS*. Chicago. IL.

Linacre, J. M. 2002. Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85–106

Linn, R.L., & Miller, M.D. 2005. *Measurement and assessment in teaching*. New Jersey: Pearson Education.

Mahmud, J. 2017. Item response theory: A basic concept. *Educational Research and Reviews*, *12*(5), 258-266. https://doi.org/10.5897/ERR2017.3147

Mashlahah, A.U. 2018. Penerapan kurikulum mengacu KKNI dan implikasinya terhadap kualitas pendidikan di PTKIN. *Edukasia: Jurnal Pendidikan Islam, 13*(1). 227-248.

Meijer, R.R., & Tendeiro, J.N. 2018. Unidimensional item response theory. In P. Irwing, T. Booth, & D. J. Hugh (Eds.), *The Wiley handbook of psychometric testing : A multidisciplinary reference on survey, scale and test development* (pp. 413-433). Wiley. https://doi.org/10.1002/9781118489772.ch15

Minbashian, A., Huon, G.F., & Bird, K.D. 2004. Approaches to studying and academic performance in short-essay exams. *Higher Education, 47*, 161–176.

Nilson, L. (2017) *Teaching at its best: A research-based resource for college instructors (4th ed.)*. San Francisco: Jossey-Bass.

Payong, M.R. 2015. Guru sebagai pekerjaan profesional dalam konteks kerangka kualifikasi nasional indonesia (KKNI). *Jurnal Pendidikan dan Kebudayaan Missio,* 7(1), 62-69.

Peraturan Menteri Riset, Teknologi, dan Pendidikan Tinggi Nomor 55 Tahun 2017 tentang Standar Pendidikan Guru (Berita Negara Republik Indonesia Tahun 2017 Nomor 1146).

Razak. N. bin Abd.. Khairani. A.Z. bin. & Thien. L.M. 2012. Examining quality of mathemtics test items using rasch model: Preminarily analysis. *Procedia - Social and Behavioral Sciences*. *69*. 2205-2214. https://dx.doi.org/10.1016/j.sbspro.2012.12.187

ISSN 2087-166X (printed)
ISSN 2721-012X (online)
DOI: https://doi.org/10.37304/jikt.v14i1.180

Reiner, C.M., Bothell, T.W., Sudweeks, R.R., & Wood, B. 2002. *Preparing effective essay questions: A Self-directed workbook for educators*. Stillwater, OK: New Forums Press.

Reynolds, C.R., Livingston, R.B., & Wilson, V.L. 2006. *Measurement and assessment in education*. Boston: Pearson.

Salend, S.J. 2011. Creating Student-Friendly Tests. *Educational Leadership, 69*(3), 52-58.

Wahyuni, L.D., Gumela, G., & Maulana, H. 2020. Interrater reliability: Comparison of essay's tests and scoring rubrics. *Journal of Physics: Conference Series*, 1933, 1-6. https://doi.org/10.1088/1742-6596/1933/1/012081

Wiggins, G. 2011. A true test: Toward a more authentic and equitable assessment. *Phi Delta Kappa*, *92*(7), 81–93.

Zile-Tamsen, C.V. 2017. Using rasch analysis to inform rating scale development. *Res High Educ, 58,* 922-933. https://doi.org/10.1007/s11162-017-9448-0