JURNAL TARBIYAH

E-ISSN : 2597-4270 | P-ISSN : 0854-2627
Volume 29, Number 2, December 2022, pp. 243-256

DETEKSI BIAS GENDER PADA INSTRUMEN EVALUASI BELAJAR KIMIA DENGAN METODE MANTEL-HAENZEL

Rizki Nor Amelia[1], Sri Rejeki Dwi Astuti[2], Anggi Ristiyana Puspita Sari[3]

[1] Universitas Negeri Semarang, Semarang, Indonesia
[2] Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[3] Universitas Palangka Raya, Palangkaraya, Indonesia
Email : [1] rizkinoramelia@mail.unnes.ac.id, [2] srirejeki.dwiastuti@yahoo.com, [3] anggi.ristiyana@fkip.upr.ac.id

DOI : http://dx.doi.org/10.30829/tar.v29i2.1781

ARTICLE INFO

Article History
Received : September 16, 2022
Reviewed : November 30, 2022
Accepted : December 15, 2022

Keywords
Item bias,
Mantel-Haenszel,
Rasch Model,
Psychometric characteristics,
Chemistry achievement

ABSTRACT

This study aims to explore psychometric characteristics and detect whether the teacher-made chemistry learning evaluation instrument contains item bias that can benefit students of a certain gender. The research sample was 358 students of class XII SMA in Yogyakarta who were taken by cluster random sampling then the answer were analyzed by Winsteps Rasch Software 3.73. In general, it can be concluded that the teacher-made chemistry learning achievement instrument is good in terms of the reliability coefficient, the distribution of the difficulty level of the items, the majority of which are in the medium category, and only one that detected as misfit item. The results of the analysis using the Mantel-Haenszel method showed that three item was biased and two of them favored male students. With the detection of bias item, it indicates that follow-up from the teacher is needed to improve these items so that the principle of test fairness can be enforced.

Pendahuluan

Pengukuran dalam bidang pendidikan (tidak terkecuali pada bidang pendidikan kimia) memang tidak mudah untuk memperoleh data yang handal. Hal ini disebabkan oleh berbagai sumber kesalahan pengukuran yang umumnya dapat dikontrol atau dikelola sebelum pengujian dilakukan (Bassey, Ovat, & Ofem, 2019). Secara umum, kesalahan pengukuran dikonseptualisasikan sebagai perbedaan antara skor aktual yang diperoleh siswa dengan skor teoretisnya (Gardner, 2013). Kesalahan pengukuran dalam penilaian pendidikan dapat berupa

kesalahan acak (tidak sistematis) atau sistematis. Dalam beberapa hal, kesalahan acak lebih mudah ditangani dan tidak terlalu serius daripada kesalahan sistematis, karena kesalahan acak umumnya mudah dideteksi, mudah diperkirakan, dan dapat dikendalikan sampai batas tertentu (Yashim, Mhab, & Waziri, 2021). Sementara itu, kesalahan sistematis merupakan kecacatan dalam skema pengukuran yang dapat bersumber dari instrumen ukur, kondisi pengujian yang tidak tepat, maupun subjek ukur; sehingga dapat berimplikasi pada ketidakakuratan skor yang diperoleh individu (Bassey, Ovat, & Ofem, 2019).

Bias butir merupakan salah satu komponen kesalahan sistematis pengukuran yang ditimbulkan akibat penggunaan instrumen pengukuran, oleh sebab itu dalam upaya pengembangan tes, guru sebaiknya memeriksa apakah butir-butir penyusun instrumen terjangkit bias atau tidak (Uyar, Kelecioglu, & Dogan, 2017). Bias butir atau lebih populer dikenal dengan nama *Differential Item Functioning* (DIF) merupakan suatu kondisi manakala *testee* dengan kemampuan yang sama namun berasal dari kelompok yang berbeda memiliki probabilitas yang berbeda dalam merespon suatu butir dengan benar (Osadebe & Agbure, 2019). Pada praktiknya, DIF digunakan sebagai salah satu indikator bahwa dalam tes yang diselenggarakan, kemungkinan terjadinya bias tes itu ada (Karami, Nodoushan, & Ali, 2011). Bias dari suatu tes merupakan suatu kondisi tes yang tidak adil (*unfair*) (Karami, 2012; Moghadam & Nasirzadeh, 2020). Apabila suatu tes memuat bias, maka penggunaan tes tersebut justru menghadirkan konsekuensi yang tidak diinginkan (Emaikwu, 2012).

Pada dasarnya, pendeteksian bias butir melibatkan evaluasi dari hubungan bersyarat antara respons butir dengan keanggotaan kelompok. Oleh sebab itu, kelompok yang dipilih seharusnya didasarkan pada pertimbangan teoritis yang mencakup apakah konstruk yang dipelajari itu dihipotesiskan memiliki konseptual yang sama atau tidak bagi lintas kelompok (Teresi, Ramirez, Lai, & Silver, 2008). Untuk melakukan deteksi DIF, kelompok yang diselidiki apakah ada butir yang bias padanya disebut kelompok fokus (*focal group*) dan kelompok pembandingnya disebut kelompok acuan (*reference group*). Kelompok Fokus atau kelompok minoritas adalah kelompok yang berpotensi dirugikan, sedangkan Kelompok Acuan adalah kelompok pembandingnya yang berpotensi diuntungkan (Karami, 2012). Dasar pengelompokan ini misalnya ditinjau dari ras, gender, agama, maupun status sosial ekonomi (Ibrahim, 2018). Dalam perspektif gender, misalnya, kelompok perempuan dapat ditentukan sebagai kelompok fokus dan kelompok acuannya adalah kelompok laki-laki; atau sebaliknya (Budiono, 2009).

Bias butir dapat terjadi dalam dua cara, yakni *uniform* DIF dan *non-uniform* DIF. Dalam *uniform* DIF, probabilitas menjawab benar suatu butir untuk satu kelompok secara konsisten lebih rendah dibandingkan kelompok yang lain (Fidelis, 2018). Hal tersebut ditunjukkan dengan adanya dua buah *Item Characteristic Curve* (ICC) yang paralel atau sejajar satu sama lain (Uyar, Kelecioglu, & Dogan, 2017), sehingga mengindikasikan tidak adanya interaksi antara tingkat kemampuan dan keanggotaan kelompok (Rustam, Naga, & Supriyati, 2019). Sementara itu, *non-uniform* DIF didefinisikan sebagai fungsi butir yang mendukung satu kelompok di beberapa tingkat kemampuan dan mendukung kelompok lain di tingkat kemampuan lain di seluruh skala kemampuan (Uyar, Kelecioglu, & Dogan, 2017). Hal tersebut ditunjukkan dengan adanya dua buah ICC yang tidak seragam (Salehi & Tayebi, 2012) atau saling berpotongan satu sama lain (Uyar, Kelecioglu, & Dogan, 2017), sehingga mengindikasikan adanya interaksi antara tingkat kemampuan dan keanggotaan kelompok (Rustam, Naga, & Supriyati, 2019) dan itu cukup sulit ditafsirkan (Salehi & Tayebi, 2012). Dalam non-*uniform* DIF, parameter daya beda butir dan tingkat kesulitan butir juga memiliki nilai yang berbeda untuk kelompok referensi dan kelompok fokus (Cuevas & Cervantes, 2012).

Penelitian ini mendasarkan bias butir pada faktor gender karena gender merupakan salah satu faktor yang umum dikenal sebagai penyebab tidak relevannya sumber varians konstruk (Bordbar, 2020). Gender sendiri merujuk pada peran, perilaku, dan identitas yang dibangun secara sosial, yang mempengaruhi bagaimana cara orang memandang diri mereka sendiri / satu sama lain, hingga bagaimana harus berperilaku dan berinteraksi dalam masyarakat (Heidari, Babor, Castro, Tort, & Curno, 2016). Pada tingkat butir, bias gender dapat diselidiki menggunakan pendekatan Teori Respons Butir (Wetzel, Hell, & Passler, 2012). Teori Respon Butir (*Item Response Theory*, IRT) berisi seperangkat model psikometrik yang digunakan dalam pengembangan, penilaian, perbaikan, penilaian, dan evaluasi (Aune, Abal, Attorresi, 2020) yang berfokus pada level butir dan skala secara keseluruhan (Toland, 2014). Kehadiran IRT sebagai sistem pengukuran yang baru, bertujuan untuk mengatasi berbagai keterbatasan yang ditemukan pada sistem pengukuran klasik atau *Classical Test Theory* (CTT) (Paek & Cole, 2019).

Model Rasch merupakan salah satu bentuk khusus dari IRT yang mempertimbangkan kemampuan responden dalam menjawab tes hanya terhadap tingkat kesukaran dari butir-butir penyusun instrumen tersebut (Azizah, Suseno, & Hayat, 2021). Model ini sangat populer digunakan untuk mendeteksi bias butir (Ukanda, Othuon, Agak, & Oleche, 2017), utamanya yang terkait dengan *uniform* DIF menggunakan metode Mantel-Haenszel. Metode Mantel-

Haenszel berbasis non-parametrik, dimana tidak perlu asumsi spesifik pada *Item Response Function* (IRF) maupun distribusi *Latent Trait* yang mendasari, sehingga mempermudah dalam perhitungannya (Li, 2015). Sejak diperkenalkan metode Mantel-Haenszel dalam penelitian DIF, telah banyak penelitian yang menguji kekuatan metode ini dan secara umum hasil penelitian tersebut menunjukkan bahwa metode Mantel-Haenszel akan memberikan hasil identifikasi DIF yang powerful manakala sampel uji yang terlibat semakin besar, *effect size* yang digunakan semakin besar, hubungan integral antara IRF dengan distribusi *latent trait* semakin tinggi, dan digunakannya Type 1 error-α dalam pengujian hipotesisnya (Li, 2015).

Penelitian bias butir (terutama dalam bidang pendidikan kimia) penting dilakukan mengingat bias butir tidak hanya merupakan komponen kunci dalam evaluasi dan validitas tes, tetapi juga sebagai komponen integral pada berbagi studi tentang keadilan tes (Ibrahim, 2018). Berdasarkan hasil penelusuran yang dilakukan, penelitian DIF yang spesifik memfokuskan pada instrumen evaluasi belajar kimia masih sangat terbatas, terutama yang menggunakan subjek pelajar Indonesia. Oleh sebab itu, penelitian ini dilakukan untuk mendeteksi ada tidaknya bias butir dalam instrumen evaluasi belajar kimia buatan guru.

Metode Penelitian

Pendekatan penelitian

Pendekatan yang digunakan dalam penelitian ini adalah pendekatan kuantitatif, dimana data penelitian didapatkan melalui teknik dokumentasi yang berupa respon jawaban siswa terhadap instrumen evaluasi belajar kimia buatan guru.

Instrumen dan subjek penelitian

Instrumen yang digunakan adalah instrumen evaluasi belajar kimia buatan guru yang berbentuk *multiple choice* dan dikerjakan oleh 358 siswa kelas XII SMA Negeri di Kota Yogyakarta (N perempuan = 206 orang, N laki-laki = 152 orang, M usia = 16,7 tahun). Melalui teknik cluster random sampling diperoleh tiga SMA Negeri, sehingga siswa kelas XII IPA dari sekolah terpilih tersebut seluruhnya menjadi sampel dalam penelitian ini. Item-item penyusun instrumen evaluasi belajar kimia yang berjumlah 40 butir berfungsi untuk mengukur kemampuan kognitif siswa yang memuat 72,5% pemahaman, 17,5% penerapan, dan 10% penalaran. Adapun ranah materi kimia yang diuji meliputi Kimia Dasar (struktur atom, sistem periodik unsur, ikatan kimia, tata nama senyawa anorganik dan organik, persamaan reaksi sederhana, dan hukum-hukum dasar kimia), Kimia Analisis (larutan (non)-elektrolit, asam-

basa, stoikiometri larutan, larutan penyangga, hidrolisis garam, kelarutan dan hasil kali kelarutan), Kimia Fisik (termokimia, laju reaksi, kesetimbangan kimia, ikatan kimia koloid, dan sifat koligatif larutan, reaksi redoks dan elektrokimia), Kimia Organik (senyawa karbon, minyak bumi, dan makromolekul: polimer, karbohidrat dan protein, serta cara analisis kuantitatifnya, lemak-minyak), dan kimia anorganik (ikatan kimia, unsur kimia yang terdapat di alam termasuk radioaktif, sifatnya, manfaatnya, kereaktifannya, dan produksinya).

Teknik analisis data

Untuk mempermudah perhitungan, software *Winsteps Rasch software* versi 3.73 (Linacre, 2022) digunakan untuk menganalisis data respon dari subyek uji. Informasi yang didapatkan dari analisis tersebut yakni identifikasi ada tidaknya butir *misfit* dan butir bias, serta mengetahui karakteristik psikometrik berupa koefisien reliabilitas yang dihasilkan dan tingkat kesulitan (*b*) bagi masing-masing butir. Dalam Model Rasch, bias butir diidentifikasi menggunakan Metode Mantel-Haenzel. Secara konvensional, deteksi DIF dimulai dengan cara mengelompokkan respons/jawaban siswa berdasarkan perbedaan gender, selanjutnya dibuat tabel kontingensi 2x2 seperti Tabel 1.

Tabel 1. Tabel Kontingensi Perhitungan DIF dengan Metode Mantel-Haenszel

	Respon		Total
	Benar (Y=1)	Salah (Y=0)	
Kelompok Acuan (R)	A_j	B_j	n_{Rj}
Kelompok Fokus (F)	C_j	D_j	n_{Fj}
	m_{1j}	m_{0j}	N_j

Lalu dilakukan analisis keberadaan DIF dengan metode Mantel-Haenszel yang berdistribusi c*hi-square* dengan menggunakan Formula 1 (Mantel-Haesnzel, 1959).

$$MH_{\chi^2} = \frac{(|\sum_j Aj - \sum_j E_{Aj}| - 0{,}5)^2}{\sum_j V_{Aj}} \quad \dots (1)$$

dimana:

$$E_{Aj} = \frac{(A_j + B_j)(A_j + C_j)}{N_j} \dots (2) \qquad \text{dan} \qquad V_{Aj} = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{N_j^2 (N_j - 1)} \dots (3)$$

Adapun hipotesis nol dan hipotesis alternatif dalam penelitian ini dirumuskan sebagai berikut:

H_0: pPj = pLj

proporsi peserta tes yang menjawab benar butir j dari kelompok perempuan (pPj) sama dengan proporsi peserta tes yang menjawab benar butir j dari kelompok laki-laki (pLj)

H_a : pPj $\neq$ pLj

proporsi peserta tes yang menjawab benar butir j dari kelompok perempuan (pPj) tidak sama dengan proporsi peserta tes yang menjawab benar butir j dari kelompok laki-laki (pLj)

Suatu butir terbukti memuat DIF (menguntungkan kelompok tertentu) manakala $MH_{\chi^2} \geq \chi^2$ tabel. Harga χ^2 tabel dapat dilihat pada tabel χ^2 pada derajat bebas=1 dengan taraf signifikansi tertentu (dalam penelitian ini $\alpha = 0.05$; sehingga χ^2 tabel = 3.841), atau dengan menginterpretasikan probabilitasnya yakni $p \leq 0.05$. Semakin besar probabilitasnya, maka semakin kecil kecenderungan butir terjangkit DIF. Ini berarti Hipotesis alternatif (H_a) ditolak atau Hipotesis nol (H_0) diterima.

Hasil

Pada bagian ini disajikan hasil penelitian yang dipaparkan secara ringkas dalam Tabel 2 yang memuat informasi tentang statistik outfit MNSQ (*outlier-sensitive or information-weighted fit Mean Square*) yang digunakan untuk mengindentifikasi apakah suatu butir termasuk *fit* atau *misfit* terhadap Model Rasch, karakteristik parameter butir berupa tingkat kesukaran butir (*b*) bagi masing-masing butir penyusun instrumen evaluasi belajar kimia buatan guru, *DIF measure* (setara dengan *b*) yang diestimasikan terpisah berdasarkan kategorisasi gendernya, serta statistik chi-square dan probabilitas yang dihasilkan dari analisis DIF menggunakan metode Mantel-Haesnzel. Berdasarkan Tabel 2 terlihat bahwa nilai statistik outfit MNSQ berkisar antara 0.77 sampai 1.88; nilai tingkat kesukaran butir berkisar antara -2.46 logit sampai 3.07 logit; *DIF measure* siswa perempuan antara -2.59 logit sampai 3.10 logit; *DIF measure* siswa laki-laki antara -2.53 logit sampai 3.04 logit; serta nilai statistik chi square dan probabilitas Mantel-Haenzel berturut-turut berkisar antara 0 sampai 4.44 dan 0.03 sampai 1.00.

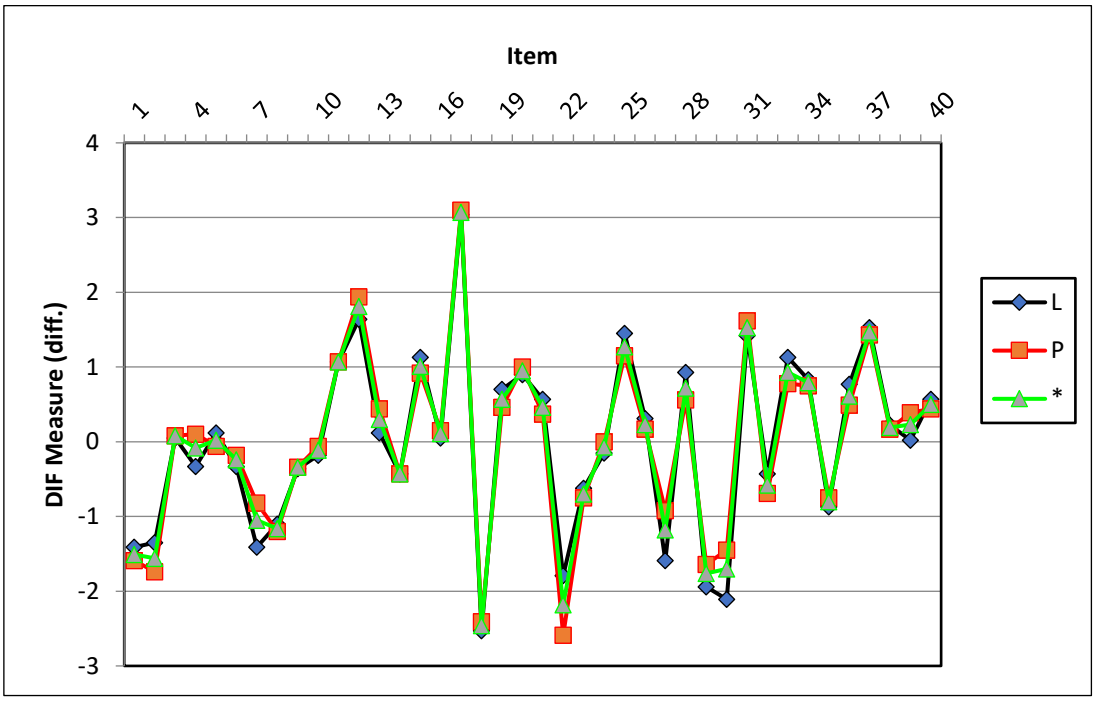
Tabel 2. Ringkasan Statistik Hasil Analisis

No	MNSQ	b	MH_{χ^2}	DIF Measure L	DIF Measure P	Probabilitas
1	1.01	-1.51	0.39	-1.41	-1.59	0.58
2	1.13	-1.56	0.01	-1.35	-1.74	0.24
3	0.97	0.08	0.17	0.05	0.08	0.92
4	1.13	-0.08	3.65	-0.33	0.10	0.09

5	0.91	0.02	0.33	0.12	-0.06	0.47
6	1.15	-0.24	0.00	-0.33	-0.18	0.57
7	1.32	-1.05	4.44	-1.41	-0.82	0.05
8	0.77	-1.16	0.04	-1.10	-1.20	0.72
9	0.92	-0.34	0.00	-0.36	-0.34	0.92
10	0.84	-0.11	0.26	-0.18	-0.06	0.60
11	0.97	1.07	0.00	1.07	1.07	1.00
12	1.03	1.81	0.35	1.64	1.94	0.27
13	0.86	0.30	1.20	0.12	0.44	0.18
14	0.87	-0.43	0.02	-.043	-0.43	1.00
15	0.99	1.01	0.34	1.13	0.92	0.38
16	1.08	0.11	0.02	0.05	0.15	0.70
17	1.88	3.07	0.26	3.04	3.10	0.86
18	0.79	-2.46	0.07	-2.53	-2.41	0.79
19	0.91	0.57	1.03	0.70	0.46	0.32
20	0.92	0.95	0.07	0.90	1.00	0.68
21	0.79	0.46	0.39	0.57	0.37	0.39
22	0.81	-2.18	2.31	-1.79	-2.59	0.05
23	1.07	-0.70	0.12	-0.62	-0.75	0.62
24	1.05	-0.07	0.48	-0.15	0.00	0.55
25	1.04	1.27	1.88	1.45	1.15	0.22
26	1.18	0.23	0.08	0.31	0.17	0.55
27	1.32	-1.18	2.95	-1.59	-0.92	0.03
28	1.21	0.72	1.15	0.93	0.56	0.12
29	1.10	-1.76	1.00	-1.94	-1.64	0.40
30	0.95	-1.70	2.71	-2.11	-1.45	0.07
31	0.78	1.53	0.74	1.42	1.62	0.43
32	0.89	-0.58	1.39	-0.43	-0.69	0.32
33	0.94	0.93	2.76	1.13	0.78	0.14
34	0.93	0.79	0.01	0.83	0.75	0.74
35	0.80	-0.80	0.08	-0.87	-0.75	0.67
36	1.07	0.61	1.75	0.77	0.49	0.24
37	1.19	1.47	0.23	1.53	1.43	0.69
38	0.96	0.19	0.01	0.22	0.17	0.85
39	1.30	0.23	3.62	0.02	0.39	0.12
40	0.97	0.50	0.26	0.57	0.44	0.57

Selain disajikan secara statistik, analisis DIF berbantuan Model Rasch ini juga memvisualisasikan data dalam bentuk plot yang menggambarkan interaksi antara butir dengan statistik biasnya. Dari Gambar 1, nampak jika sebagian besar butir yang direspon oleh subjek ukur sudah memiliki kurva yang puncaknya hampir tumpang tindih. Ini artinya, pada butir-butir

tersebut tidak terjangkit bias. Pada butir yang terjangkit bias dengan statistik yang signifikan, puncak kurva akan nampak memiliki jarak, dimana posisi butir yang direspon siswa laki-laki bisa saja terletak diatas butir yang direspon siswa perempuan, atau sebaliknya; sebagaimana yang terjadi pada butir nomor 7, 22, dan 27. Plot ini sekaligus menjadi penjelas bagi nilai probabilitas, sehingga dapat ditentukan kelompok mana yang diuntungkan maupun dirugikan oleh butir.



Gambar 1. Person DIF Plot

Pembahasan

Sebagaimana yang telah dijelaskan dalam pendahuluan. penelitian ini berfokus untuk mengidentifikasi uniform DIF yang didasakan pada faktor gender menggunakan Metode Mantel-Haenszel berbantuan Model Rasch. Secara matematis, Model Rasch setara dengan model 1-PL (1-Parameter Logistik) dalam IRT yang memiliki asumsi bahwa daya diskriminasi dan indeks tebakan semu pada masing-masing butir diabaikan atau dianggap konstan (Humpry, 2015) dan nilai dari satu model dapat ditransformasikan ke model lainnya dengan penskalaan ulang yang sesuai (Hayat, Putra, & Suryadi, 2020). Sebelum melakukan deteksi bias butir, terlebih dahulu dilakukan identifikasi guna menilai kecocokan data (*item fit*) terhadap Model Rasch yang digunakan. Melalui *item fit*, dapat dijelaskan sejauhmana pola sampel respon terhadap suatu butir itu konsisten seperti respon orang lain dalam menanggapi butir-butir yang

lain (Razak, Khairani, & Thien, 2012). Adapun ukuran statistik yang digunakan untuk kriteria *item fit* adalah Outfit-MNSQ, dimana suatu butir dikatakan fit manakala berada pada rentang MNSQ yang ditentukan. yakni $0.5 \leq \text{MNSQ} \leq 1.5$ (Linacre, 2022). Berdasarkan kriteria tersebut, terlihat bahwa terdapat satu butir yang teridentifikasi sebagai item misfit, yakni butir nomor 17. Hadirnya item misfit layak dievaluasi apakah perlu dipertahankan dalam model pengukuran dengan cara membandingkan koefisien reliabilitas ketika item misfit tersebut disertakan dalam tes atau tidak. Penurunan keandalan setelah penghapusan item misfit dapat dianggap sebagai alasan untuk mempertahankannya meskipun ada ketidaksesuaian (Kohler & Hartig, 2017).

Terkait reliabilitas, instrumen evaluasi belajar kimia buatan guru telah memiliki keandalan dalam kategori baik, yang ditunjukkan melalui koefisien *model person reliability* dan *real person reliability* berturut-turut adalah 0.86 dan 0.85; sedangkan koefisien *item model reliability* dan *item real reliability* keduanya 0.99. Sementara itu, untuk tingkat kesukaran butir, Hambleton & Swaminathan (1985) menyebutkan bahwa nilai *b* yang mendekati -2.0 logit berarti butirnya mudah sedangkan nilai *b* yang mendekati +2.0 logit berarti butirnya sukar. Hasil analisis menunjukkan bahwa sebagian besar (92.5%) tergolong butir dengan tingkat kesukaran sedang. 5% butir tergolong mudah. dan 2.5% sisanya tergolong butir sukar.

Dalam software Winsteps yang berbasis Model Rasch, sebetulnya terdapat dua metode yang disediakan untuk melakukan deteksi keberadaan DIF, yakni Mantel-Haenszel chi-square dan Welch t-test. Secara teoritis kedua metode tersebut memang akan memberikan interpretasi yang sama (Shanmugam, 2018), namun dalam praktiknya metode Mantel-Haenszel memberikan estimasi lebih akurat karena kemampuannya dalam memprediksikan data yang hilang, sehingga metode ini lebih disukai untuk deteksi DIF dibandingkan dengan metode Welch t-test (Linacre, 2022). Penelitian ini menggunakan siswa perempuan sebagai kelompok fokus yang diselidiki apakah ada butir yang bias padanya dan kelompok siswa laki-laki sebagai kelompok referensi atau kelompok pembandingnya. Hasil analisis baik ditinjau dari statistik Mantel Haenszel chi square maupun probabilitasnya menunjukkan bahwa terdeteksi tiga butir yang terjangkit bias gender, yakni butir nomor 7, 22, dan 27 yang disajikan dalam Gambar 2, Gambar 3, dan Gambar 4.

7. Kalsium dan oksigen membentuk senyawa sebagai berikut :

No	Massa Ca (g)	Massa Oksigen (g)	Massa Kalsium Oksida (g)
1	5	6	7
2	10	4	14
3	15	10	16
4	20	14	28

Perbandingan massa kalsium : oksigen pada senyawa tersebut adalah ....

A. 2 : 3
B. 2 : 5
C. 3 : 2
D. 5 : 2
E. 5 : 3

Gambar 2. Butir Soal Nomor 7

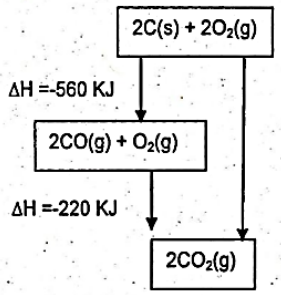
22. Diketahui senyawa turunan benzena :
1. Toluena
2. Anilin
3. Fenol
4. Asam benzoat
5. Stirena

Senyawa yang dapat digunakan untuk membuat TNT dan senyawa untuk pengawet makanan berturut- turut adalah ....

A. 1 dan 2
B. 1 dan 4
C. 2 dan 3
D. 3 dan 4
E. 4 dan 5

Gambar 3. Butir Soal Nomor 22

27. Perhatian diagram energi berikut!



Dari diagram tersebut ,besarnya perubahan entalpi pembentukan CO_2 adalah ....

A. –760 KJ/mol
B. –390 KJ/mol
C. –190 KJ/mol
D. +390 KJ/mol
E. +760 KJ/mol

Gambar 4. Butir Soal Nomor 27

Berdasarkan Gambar 1, nampak jika butir nomor 7 dan 27 lebih menguntungkan siswa laki-laki dibanding siswa perempuan karena kurva DIF siswa perempuan (*DIF measure* butir 7 = -0.82 ; *DIF measure* butir 27 = -1.79) terletak lebih diatas kurva DIF siswa laki-laki (*DIF measure* butir 7 = -1.41; DIF measure butir 27 = -2.59). Sementara itu, butir nomor 22 lebih menguntungkan siswa perempuan dibanding siswa laki-laki karena kurva DIF siswa laki-laki (*DIF measure* = -1.79) terletak lebih diatas kurva DIF siswa perempuan (*DIF measure* = -2.59). Untuk menjawab butir 7 dan 27, diperlukan kemampuan kuantitatif yang kuat. Misalnya pada

nomor 7, terlebih dahulu siswa harus mencari nilai yang sama antara massa zat dari reaktan mula-mula dan massa zat produk hasil reaksi, baru kemudian membuat perbandingnya. Untuk nomor 27, siswa harus membuat persamaan matematis berdasarkan diagram entalpinya, baru kemudian menghitung besarnya entalpi pembentukan dalam satu mol CO_2. Sementara pada butir 22, siswa dituntut untuk memahami struktur dari produk senyawa yang ditanyakan, baru kemudian menghubungkannya dengan senyawa turunan benzena yang digunakan untuk membuatnya.

Deteksi bias butir diperlukan tidak hanya untuk memastikan bahwa butir secara konten telah sesuai, namun juga membantu *test developers* (dalam hal ini guru kimia) untuk mengidentifikasi butir yang berfungsi secara berbeda pada dua kelompok yang diuji karena kriteria kesetaraan butir seharusnya berasal dari hasil analisis respon peserta tes itu sendiri (Ibrahim, 2018). Pada akhirnya, temuan penelitian ini mendukung temuan lainnya yang menyiratkan bahwa instrumen evaluasi belajar kimia memang tidak luput dari bias butir (e.g. Queensoap & Orluwene, 2019; Kendhammer, Holme, & Murphy, 2013; Fidelis, 2018). Selain itu, berbagai hasil penelitian juga menunjukkan jika variabel gender memang berpengaruh terhadap kemampuan seseorang dalam mempelajari kimia (direfleksikan dengan prestasi belajar yang diraih), dimana secara spesifik disebutkan jika siswa laki-laki memang lebih unggul dibandingkan siswa perempuan (e.g. Amunga, Amandalo, & Musera, 2011; Ezeudu & Obi-Theresa, 2013; Tenaw, 2013; Veloo, Hong, & Lee, 2015). Keunggulan ini secara umum menjadikan banyak negara khawatir tentang jumlah perempuan yang menekuni bidang sains ke depannya (D' Andola, 2016). Studi yang dilakukan Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman (2012) memberikan fakta bahwa pelamar laki-laki dinilai lebih kompeten daripada pelamar perempuan untuk posisi manajer laboratorium maupun sebagai pengajar di Jurusan Biologi, Kimia, dan Fisika. Meskipun ilmu kimia memiliki karakteristik yang cenderung abstrak dan memerlukan ketrampilan matematis terutama dalam menyelesaikan berbagai persamaan kimia (Iwuanyanwu, 2021) yang berarti membutuhkan beberapa keahlian tambahan, namun itu bukan alasan sehingga berimbas menguntungkan atau merugikan gender tententu baik dalam pembelajaran maupun pilihan karir ke depan.

Kesimpulan, Implikasi, dan Saran

Berdasarkan analisis yang telah dilakukan, dapat disimpulkan secara umum bahwa instrumen evaluasi belajar kimia buatan guru memiliki karakteristik psikometrik yang baik ditinjau dari koefisien keandalan yang baik dengan distribusi tingkat kesukaran butir mayoritas

pada kategori sedang. Ditinjau dari kecocokan model yang menggambarkan validitas pengukuran model Rasch. terdeteksi satu *misfit item*. Sementara deteksi bias butir dengan Metode Mantel-Haenszel menunjukkan terdapat tiga butir terjangkit bias gender. Terkait bias, ada kecenderungan siswa laki-laki lebih diuntungkan dalam mengerjakan butir yang melibatkan kemampuan kuantitatif, sedangkan siswa perempuan lebih mungkin diuntungkan dalam mengerjakan butir yang melibatkan kemampuan verbal. Hadirnya butir bias dapat menjadi bahan reflektif bagi guru untuk memperbaiki butir-butir tersebut. Selain itu, perlu pula diberikan penguatan kemampuan kuantitatif bagi siswa perempuan dan kemampuan verbal bagi siswa laki-laki. Pada akhirnya, melalui deteksi bias butir, akan diperoleh manfaat berupa terciptanya pengujian yang adil, minimalnya ancaman terhadap validitas internal, dan pemahaman yang terkait dengan respon butir terhadap performansi tes untuk melihat apakah proses tersebut berlaku sama pada individu yang berasal dari kelompok yang berbeda.

Daftar Pustaka

Amunga. J. K.. Amadalo. M. M.. & Musera. G. (2011). Disparities in chemistry and biology achievement in secondary schools: Implications for vision 2030. *International Journal of Humanities and Social Science*. *1*(18). 226-236.

Aune, S. E., Abal, F. J. P., & Attorresi, H. F. (2020). A psychometric analysis from the Item Response Theory: step-by-step modelling of a loneliness scale. *Ciencias Psicologicas*, *14*(1), 1-15. https://dx.doi.org/10.22235/cp.v14i1.2179

Azizah, N., Suseno, M., Hayat, B. (2021). Item analysis of rasch model items in the final semester exam indonesian language lesson. *World Journal of English Language*, *12*(1), 15-26. URL: https://doi.org/10.5430/wjel.v12n1p15

Bassey, B. A, Ovat, S. V., & Ofem, U. J. (2019). Systematic error in measurement: Ethical implication in decision makin in learners' assessment in the Nigerian educational system. *Prestige Journal of Education*, *2*(1), 137-146.

Bordbar, S. (2020). Gender differential item functioning (GDIF) analysis in Iran's University Entrance Exam. *English Language in Focus*, *3*(1), 49-68.

Budiono. (2009). The accuracy of mantel-haenszel. sibtest. and regression methods in differential item function detection. *Jurnal Penelitian dan Evaluasi Pendidikan*. *12(*1). 1-20.

Cuevas, M., & Cervantes, V. H. (2012). Differential item functioning detection with logistic regression. *Mathematics and Social Sciences*. *3*. 45-59.

D'Andola, C. (2016). Women in chemistry-where we are today. *Chemistry European Journal*, *22*, 3523-3528. https://dx.doi.org/10.1002/chem.201600474

Emaikwu. S. O. (2012). Issues in test item bias in public examinations in Nigeria and implications for testing. *International Journal of Academic Research in Progressive Education and Development*. *1*(1). 175-187.

Ezeudu. F. O. & Obi-Theresa. N. (2013). Effect of gender and location on students' achievement in chemistry in secondary schools in Nsukka local government area of Enugu state Nigeria. *Research on Humanities and Social Sciences*. *3*(15). 50-55.

Fidelis. I. (2018). Use of Differential Item Functioning (DIF) analysis for bias analysis in test construction. *International Journal of Education. Learning and Development*. *6*(3). 80-91.

Gardner. J. (2013). The public understanding of error in educational assessment. *Oxford Review of Education. 39*. 72–92. https://dx.doi.org/10.1080/03054985.2012.760290

Hambleton. R.K.. & Swaminathan. H. (1985). *Items response theory: Principles and application.* Boston: Kluwer-Nijjhoff Publish.

Hayat, B., Putra, M. D. K., & Suryadi, B. (2020). Comparing item parameter estimates and fit statistics of the Rasch Model from three different tradition. *Jurnal Penelitian dan Evaluasi Pendidikan*, *24*(1), 39-50. https://dx.doi.org/10.21831/pep.v24i1.29871

Heidari, S., Babor, T. F., Castro P. D.,Tort S., & Curno, M. (2016). Sex and gender equity in research: rationale for the SAGER guidelines and recommended use. *Research Integrity and Peer Review*, *1*(2), 1-9. https://dx.doi.org/10.1186/s41073-016-0007-6

Humpry, S. (2015). Using a rasch model to account for guessing as a source of low discrimination, *J App Meas*, *16*(2), 193-203.

Ibrahim. A. (2018). Differential item functioning: The state of the art. *Jigawa Journal of Multidisciplinary Studies. 1*(1). 37-50.

Iwuanyanwu, P. N. (2021). Adrresing common deficiencies of mathematics skill among chemistry student teachers. *African Journal of Educational Studies in Mathematics and Science*, *17*(1), 1-17. https://dx.doi.org/10.4314/ajesms.v17i1.1

Karami, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment*, *11*(2), 59-75.

Karami. H.. Nodoushan. M. A. S.. & Ali. M. (2011). Differential item functioning (DIF): Current problems and future directions. *International Journal of Language Studies*. *5*(3). 133-142

Kendhammer, L., Holme, T., & Murphy, K. (2013). Identifying differential performance in general chemistry: Differential item functioning Analysis of ACS General Chemistry Trial Tests. *Journal of Chemical Education*, *90*, 846-853. https://dx.doi.org/10.1021/ed4000298

Kohler, C., & Hartig, J. (2017). Practical significance of item misfit in educational assessment. *Appl Psychol Meas*, *41*(5), 388-400, https://dx.doi.org/10.1177/0146621617692978

Li. Z. (2015). A power formula for the Mantel–Haenszel test for differential item functioning. applied psychological measurement. *Applied Psychological Measurement*. *39*(5). 373–388. https:/dx.doi.org/10.1177/0146621614568805

Linacre. J.M. (2022). *A user's guide to WINSTEPS*. Chicago. IL: Winsteps

Moghadam, M., & Nasirzadeh, F. (2020). The application of Kunnan's test fairness framework (TFF) on a reading comprehension test. *Language Testing in Asia*, *10*(7), 1-21. https://doi.org/10.1186/s40468-020-00105-2

Moss-Racusin, C.A., Dovidio, J.F., Brescoll, V.L., Graham, M.J., & Handelsman, J. (2012). Science faculty´s subtle gender biases favor male students. *Proceedings of National Academy of Sciences of the United States of America*, *109*(41), 16474-16479. https://dx.doi.org/10.1073/pnas.121128610

Osadebe, P.U., & Agbure, B. (2019). Assessment of differential item functioning in social studies multiple choice questions in basic education certificate examination. *European Journal of Education Studie*s, *6*(8), 312-344. https://dx.doi.org/10.5281/zenodo.3674732

Paek, I., & Cole, K. (2019). *Using R for Item Response Theory Model Applications*. New York, NY: Routledge.

Queensoap, M., & Orluwene, G.W. (2019). Use of mantel haenszel differential item functioning in detecting item bias in a chemistry achievement test in four ethnic groups in Nigeria.

International Journal of Current Research, 11(3), 2665-2670. https://dx.doi.org/10.24941/ijcr.34709.03.2019

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

Rustam, A. Naga, D.S., & Supriyati, Y. (2019). A comparison of mantel-haenszel and standardization methods: Detecting differential item functioning. Jurnal Matematika dan Pembelajaran, 7(1), 16-31. https://dx.doi.org/10.24252/mapan.2019v7n1a2

Salehi, M., & Tayebi, A. (2012). Differential item functioning: Implications for test validation. *Journal of language teaching and research, 3*(1), 84-95. https://dx.doi.org/10.4304/jltr.3.1.84-92

Shanmugam. K.S. (2018). Determining gender differential item functioning for mathematic in coeducational school culture. *Malaysian Journal of Learning and Instruction*. *15*(2). 83-109.

Tenaw. Y. A. (2013). Relationship between self-efficacy. academic achievement and gender in analytical chemistry at Debre Markos College of teacher education. *African Journal of Chemical Education*. *3*(1). 3-28.

Toland, M.D. (2014). Practical guide to conducting an item response theory analysis. *Journal of Early Adolescence*. *34*(1), 120-151. https://dx.doi.org/10.1177/0272431613511332

Ukanda, F., Othuon, L., Agak, J., Oleche, P. (2017). Effect of sample size, ability distribution, and test length on detection of differential item functioning using mantel-haenszel statistic. *International Journal of Education and Research, 5*(5), 91-104.

Uyar. S.. Kelecioglu. H.. & Dogan. N. (2017). Comparing differential item functioning based on manifest groups and latent classes. *Educational science: Theory and practice*. *17*(6). 1977-2000. https:/dx.doi.org/10.12738/estp.2017.6.0526

Veloo. A.. Hong. L.H.. & Lee. S.C. (2015). Gender and ethnicity differences manifested in chemistry achievement and self-regulated learning. *International Education Studies*. *8*(8). http://dx.doi.org/10.5539/ies.v8n8p1Hayat, Putra, & Suryadi, 2020

Wetzel. E.. Hell.. B.. & Passler. K. (2012). Comparison of different test construction strategies in the Wright. B.. & Stone. M. (1999). *Measurement essentials (2nd ed.)*. Wilmington: Wide Range. Inc.

Yashim, A.U., Mhab., L.C., & Waziri, J.A. (2021). Measurement errors in educational assessment. *Journal of Educational Theory and Practice*, *1*(1), 1-9.