



Four-tier Diagnostic Test sebagai Pendeteksi Miskonsepsi pada Materi Ikatan Kimia untuk Mendukung Pendidikan Berkelanjutan

Munawaroh*, Das Salirawati

Program Studi Pendidikan Kimia, Universitas Negeri Yogyakarta, Indonesia

* Korespondensi Penulis. E-mail: munawaroh2802@gmail.com

Abstrak

Pemahaman konsep kimia sangat penting untuk mencegah miskonsepsi, terutama pada materi ikatan kimia. Namun, belum banyak instrumen yang dapat secara sistematis mengidentifikasi miskonsepsi peserta didik. Penelitian ini bertujuan mengembangkan instrumen *four-tier diagnostic test* yang valid dan reliabel untuk mengidentifikasi miskonsepsi pada materi ikatan kimia. Penelitian menggunakan metode *Research and Development* (R&D) dengan model 4D (*define, design, develop, dan disseminate*). Uji coba dilakukan dalam tiga tahap meliputi uji coba ahli, uji coba skala kecil di satu sekolah, dan skala besar di 3 sekolah. Data dianalisis menggunakan uji validitas, reliabilitas, tingkat kesukaran, dan daya pembeda. Hasil menunjukkan bahwa instrumen ini valid dan reliabel dengan tingkat kesukaran dan daya pembeda yang bervariasi. Instrumen ini efektif mengidentifikasi miskonsepsi serta hubungan antara keyakinan dan pemahaman konsep peserta didik.

Kata Kunci: *Four-tier diagnostic test*, Miskonsepsi ikatan kimia, Pengembangan instrumen

Four-tier Diagnostic Test to Identify Misconceptions in Chemical Bonding Material to Support Sustainable Education

Abstract

Understanding chemical concepts is crucial for preventing misconceptions, particularly regarding chemical bonding. However, few instruments can systematically identify students' misconceptions. This study aims to develop a valid and reliable four-tier diagnostic test instrument to identify misconceptions regarding chemical bonding. The study employed the Research and Development (R&D) method with the 4D model (define, design, develop, and disseminate). The trial was conducted in three stages: expert trials, small-scale trials in one school, and large-scale trials in three schools. Data were analyzed using validity, reliability, difficulty level, and discriminatory power tests. The results showed that this instrument was valid and reliable with varying levels of difficulty and discriminatory power. This instrument effectively identified misconceptions and the relationship between students' beliefs and conceptual understanding.

Keywords: *Four-tier diagnostic, Misconceptions in chemical bonding, Instrument development*

How to Cite: Munawaroh, M., & Salirawati, D. (2025). Four-tier diagnostic test sebagai pendeteksi miskonsepsi pada materi ikatan kimia guna mendukung pendidikan berkelanjutan. *Jurnal Pendidikan Matematika dan Sains*, 13(Special Issue), 176–188. https://doi.org/10.21831/jpms.v13iSpecial_issue.88643

Permalink/DOI: DOI: https://doi.org/10.21831/jpms.v13iSpecial_issue.88643

PENDAHULUAN

Kimia merupakan salah satu cabang ilmu pengetahuan dari lingkup *science* yang mempelajari mengenai susunan, struktur, sifat, dan perubahan materi (Rahayu, 2009) (Tümay, 2016). Pada proses pembelajaran, kimia dianggap menjadi salah satu materi yang sulit dipahami peserta didik (Üce & Ceyhan, 2019) (Özmen, 2004). Hal ini disebabkan karakteristik kimia pada pembelajarannya disajikan pada

tingkat representasi yang berbeda, seperti mikroskopis (mikro), makroskopik (makro), dan simbolik (Derkach, 2021; Tümay, 2016; Härmälä-Braskén et al., 2020; Atkins, 2013; Pratama & Rohaeti, 2024). Ketiga karakteristik tersebut sama pentingnya untuk dipahami dan hal ini juga menjadi salah satu contoh bahwa pemahaman konsep kimia itu sangat penting. Permasalahannya pada proses pembelajaran peserta didik sering kali mengalami kesalahan konsep atau miskonsepsi, meskipun materi

pembelajaran tersebut sudah dijelaskan oleh pendidik (Taber, 2009).

Talib et al. (2018), mengatakan bahwa miskonsepsi dapat didefinisikan sebagai konsepsi yang bertentangan atau berbeda dengan pandangan ilmiah yang diterima saat ini. Miskonsepsi dapat terjadi disebabkan oleh beberapa faktor, seperti informasi yang didapatkan peserta didik dari internet (Sesen, 2010), pengetahuan awal peserta didik, pengalaman sehari-hari, kebingungan terminologi ilmiah, dan representasi yang salah dalam buku teks sebagai sumber (Tümay, 2016), atau bahkan miskonsepsi ini dapat disebabkan oleh pendidik (Härmälä-Braskén et al., 2020). Pendapat Üce & Ceyhan, (2019) mengatakan bahwa berbagai praktik di kelas dapat menyebabkan miskonsepsi bagi peserta didik. Hal ini juga dibenarkan oleh Taber (2009), misalnya ketika peserta didik datang ke kelas dengan konsepsi alternatif tentang suatu topik, mereka sering memahami penjelasan pendidik dikaitkan dengan pemahaman mereka sendiri. Hal ini dapat menyebabkan distorsi terhadap konsep yang dimaksudkan pendidik, sehingga memperumit pemahaman. Hal ini diperkuat dengan pernyataan Özmen (2004), bahwa keyakinan peserta didik yang sudah ada sebelumnya mempengaruhi bagaimana peserta didik mempelajari pengetahuan ilmiah baru dan memainkan peran penting dalam pembelajaran. Oleh karena itu sangat penting mengidentifikasi adanya miskonsepsi lebih awal agar pendidik dapat mengubah miskonsepsi tersebut menjadi sebuah konsep yang sesuai dan benar.

Banyak penelitian mengenai identifikasi miskonsepsi pada materi kimia, misalnya pada materi elektrokimia (Murniningsih et al., 2020), redoks (Rahmiati et al., 2022), tingkat keasaman (Tümay, 2016), anorganik (Derkach, 2021), kimia kuantum dasar (Stefani & Tsaparlis, 2009), asam basa (Talib et al., 2018), kimia kinetik (Habiddin & Page, 2019), laju reaksi (Harahap & Novita, 2020), elektrolit dan non elektrolit (Hadinugrahaningsih et al., 2020), ikatan kimia (Setiawan & Ilahi, 2022), dan lainnya.

Pada penelitian yang dilakukan oleh Coll & Taylor (2002), menyimpulkan bahwa materi ikatan kimia merupakan salah satu topik yang sering dianggap bermasalah oleh peserta didik. Selain itu, pada penelitian yang dilakukan Üce & Ceyhan (2019), menyatakan bahwa materi ikatan kimia (ikatan kovalen, ikatan ion, ikatan hidrogen dan juga geometri molekul)

merupakan materi yang dianggap abstrak, kompleks, dan juga sulit dipahami peserta didik. Pada miskonsepsi tentang ikatan, terlihat bahwa beberapa peserta didik tampak bingung mendefinisikan ikatan ion dan ikatan kovalen (Özmen, 2004).

Berdasarkan latar belakang masalah dengan didasari urgensi, tren penelitian pengembangan instrumen miskonsepsi kimia, dilakukan pengembangan instrumen *four-tier diagnostic test* pada materi ikatan kimia. Pertanyaan penelitian yang mendasari yaitu “Bagaimana karakteristik instrumen *four-tier diagnostic test* pada materi ikatan kimia yang telah dikembangkan ditinjau dari: validitas, reliabilitas, tingkat kesukaran butir soal dan daya beda soal?”

METODE

Penelitian ini merupakan *Research and Development* yang memiliki tujuan utama mengembangkan suatu produk tertentu dengan menguji keefektifan produk tersebut (Sugiyono, 2010). Penelitian ini bertujuan untuk mengembangkan instrumen *four-tier diagnostic test* untuk mengidentifikasi miskonsepsi peserta didik yang valid dan reliabel dalam pembelajaran kimia. Pengembangan ini menggunakan model pengembangan 4-D (*define, design, develop, dan disseminate*). Uji coba diterapkan di SMAN yang berada di Kecamatan Sekayu, Kabupaten Musi Banyuasin, Provinsi Sumatera Selatan.

Populasi sebanyak 191 peserta didik. Sampel penelitian ini menggunakan teknik *purposive sampling*. Karakteristik sampel penelitian meliputi siswa kelas XI yang telah mempelajari materi ikatan kimia dan memiliki kemampuan akademik yang beragam, dari rendah, sedang, hingga tinggi, agar variasi jawaban dan pemahaman tentang konsep kimia dapat terukur dengan baik.

Prosedur pengembangan pada penelitian ini mengadaptasi dari model penelitian 4D yang telah digunakan oleh Fratiwi et al. (2017). Tahapan meliputi:

Define

Pada tahap ini bertujuan untuk menetapkan dan mendefinisikan persyaratan instruksional. Hal yang dilakukan adalah melakukan analisis dasar permasalahan (*front-end analysis*) dengan menganalisis dasar permasalahan yang dihadapi pendidik pada permasalahan kelemahan belajar peserta didik

melihat dari beberapa sumber dan melakukan wawancara kepada salah satu guru. Kedua, analisis siswa (*learner analysis*) dilakukan dengan mengkaji terhadap sasaran instrumen soal yang dibuat. Ketiga, analisis tugas (*task analysis*) terkait alat evaluasi yang sudah digunakan oleh pendidik. Keempat, analisis lonsep (*concept analysis*) Analisis ini dilakukan dengan mengidentifikasi konsep-konsep utama yang akan dijadikan acuan sebagai dasar pemecahan masalah dari hasil analisis yang

Design

Tahap ini bertujuan untuk merancang desain awal instrumen soal. Hal yang dilakukan meliputi pemilihan model evaluasi (*evaluation selection model*), pemilihan format (*format selection*), dan desain awal (*initial design*). Hasil dari tahap *design* akan dilanjutkan pada tahap *develop*.

Develop

Tahapan ini meliputi validasi dan revisi instrumen *four-tier diagnostic test* untuk mendapatkan penilaian validitas teoritik dan saran perbaikan guna menghasilkan produk pengembangan yang valid dan berkualitas. Beberapa validator diminta untuk mengevaluasi dari segi materi dan *assessment*. Selanjutnya dilakukan uji coba instrumen untuk mengukur validitas empiris, reliabilitas, tingkat kesukaran, dan daya pembeda.

Disseminate

Tahap terakhir meliputi diseminasi. Pada tahap ini dilakukan pengemasan akhir dan pendistribusian instrumen kepada guru. Diharapkan instrumen ini dapat menambah variasi penilaian kimia untuk menyongsong pendidikan yang berkelanjutan.

Teknik pengumpulan data dalam penelitian ini dilakukan melalui metode tes dan non-tes. Penelitian ini melibatkan empat jenis data.

1. Data hasil wawancara yang diperoleh melalui teknik tertulis menggunakan lembar wawancara terstruktur. Wawancara dilakukan dengan guru kimia SMA/MA menggunakan instrumen panduan wawancara.
2. Data hasil validasi ahli materi yang dilakukan dengan teknik *checklist* pada lembar validasi ahli materi. Aspek yang dinilai mencakup tiga aspek penilaian yaitu

kebenaran konsep, keluasan materi, dan kedalaman materi.

3. Data hasil validasi ahli *assessment* yang dilakukan dengan teknik *checklist*. Validasi ini menggunakan lembar validasi ahli *assessment* yang menilai tiga aspek yaitu aspek materi, aspek konstruksi, dan aspek bahasa.
4. Data hasil miskonsepsi peserta didik pada ikatan kimia yang dilakukan dengan teknik tes menggunakan instrumen *four-tier diagnostic test*.

Analisis data pada instrumen *four-tier diagnostic test* sebagai berikut.

1. Uji Validitas

Setelah revisi selesai, instrumen diujikan ke peserta didik dan dilakukan validitas empiris, reliabilitas, tingkat kesukaran, daya pembeda, dan mengetahui hasil diagnosis miskonsepsi peserta didik sesuai dengan kombinasi jawaban peserta didik. Hasil analisis validitas dilihat dan dianalisis dengan program *ministep* pada tabel *Out fit order* untuk mengetahui kesesuaian butir soal yang berfungsi dalam kategori normal untuk digunakan sebagai tingkat kevalidan pengukuran miskonsepsi peserta didik. Pada analisis validasi instrumen yang digunakan hanya *outfit mean square* (MNSQ) dan *outfit Z-standard* (ZSTD) dengan memperhatikan beberapa syarat menurut Boone dalam (Sumintono & Widhiarso, 2015) pada Tabel 1.

Tabel 1. Kriteria validitas empiris

Outfit	Kategori
<i>Outfit Mean Square</i> (MNSQ)	$0,5 < \text{MNSQ} < 1,5$
<i>Outfit Z-Standard</i> (ZSTD)	$-2,0 < \text{ZSTD} < +2,0$

2. Reliabilitas

Suatu tes yang diberikan ke peserta didik selain harus valid juga harus reliabel. Analisis Rasch dapat memberikan beberapa nilai reliabilitas, yaitu *Alfa Cronbach*, *Pearson Reliability*, dan *item reliability*. *Alfa Cronbach* menunjukkan interaksi antara *person* dan semua butir soal secara utuh, *Pearson Reliability* menunjukkan konsistensi peserta didik dalam menjawab, sedangkan *item reliability* menunjukkan kualitas pada suatu item tes. Interpretasi dari *Alfa Cronbach*, *Pearson Reliability*, dan *item reliability* menurut Sumintono & Widhiarso (2015:85) tersaji pada Tabel 2, Tabel 3, dan Tabel 4.

Tabel 2. Interpretasi *Alfa Cronbach*

Nilai	Interpretasi
$\alpha < 0,5$	Sangat Buruk
$0,5 \leq \alpha \leq 0,6$	Buruk
$0,6 \leq \alpha \leq 0,7$	Cukup
$0,7 \leq \alpha \leq 0,8$	Bagus
$\alpha > 0,8$	Bagus Sekali

Tabel 3. Interpretasi *Person Reliability*

Nilai	Interpretasi
Nilai $< 0,67$	Lemah
$0,67 \leq \text{Nilai} \leq 0,80$	Cukup
$0,80 \leq \text{Nilai} \leq 0,90$	Bagus
$0,91 \leq \text{Nilai} \leq 0,94$	Bagus sekali
Nilai $> 0,94$	Istimewa

Tabel 4. Interpretasi *item reliability*

Nilai	Interpretasi
Nilai $< 0,67$	Lemah
$0,67 \leq \text{Nilai} \leq 0,80$	Cukup
$0,80 \leq \text{Nilai} \leq 0,90$	Bagus
$0,91 \leq \text{Nilai} \leq 0,94$	Bagus sekali
Nilai $> 0,94$	Istimewa

3. Tingkat kesukaran soal

Pada analisis Rasch untuk mengetahui tingkat kesukaran butir soal dapat dilihat di tabel

item measure yang kemudian dikategorikan menurut (Purniasari et al., 2021:210), dapat dilihat pada Tabel 5.

Tabel 5. Tingkat kesukaran soal

Nilai <i>Measure</i>	Kategori
< -1	Sangat mudah
$-1 \text{ sampai } 0$	Mudah
$0 \text{ sampai } 1$	Sukar
> 1	Sangat Sukar

4. Daya Beda Soal

Penelitian ini menggunakan perangkat lunak MINISTEP 5.2.3 dengan *output Table 10 item fit order* yang akan dianalisis dengan Rasch. Agar dapat mengetahui kategori daya

pembeda setiap butir instrumen dapat dilihat dari kolom *PT-Measure Corr*. Interpretasi dari *PT-Measure Corr* menurut Brown dalam (Smiley, 2015:24), dapat dilihat pada Tabel 6.

Tabel 6. Daya beda soal

<i>PT Mean Corr</i>	Interpretasi
$0,40 < ID$	Sangat baik
$0,30 < ID \leq 0,40$	Baik
$0,20 < ID \leq 0,30$	Kurang baik
$ID < 0,20$	Tidak baik

HASIL DAN PEMBAHASAN

Hasil validitas butir soal disajikan pada Tabel 7. Dilihat dari hasil validitas setiap butir soal pada Tabel 7 menunjukkan hasil seluruh soal sesuai dengan kriteria sehingga setiap butir soal dapat digunakan atau pola respon peserta didik keseluruhannya dinilai fit yang berarti

keseluruhan butir soal *four-tier diagnostic test* berfungsi secara normal pada pengukuran. Hal ini sejalan dengan pendapat Rohmah et al. (2022:120) mengatakan butir soal yang dapat berfungsi dengan baik atau tidak dalam pengukuran dilihat dari *item fit* pada setiap butir soal.

Tabel 7. Hasil validitas

No. Soal	Nilai Outfit MNSQ	Nilai Outfit ZSTD	Keterangan
1.	0,99	-0,04	Diterima
2.	1,02	0,15	Diterima
3.	1,04	0,27	Diterima
4.	0,91	-0,52	Diterima
5.	0,94	-0,23	Diterima
6.	0,77	-1,11	Diterima
7.	1,04	0,23	Diterima
8.	1,09	0,39	Diterima
9.	0,99	0,30	Diterima
10.	0,95	-0,36	Diterima
11.	1,04	0,25	Diterima
12.	1,05	0,26	Diterima
13.	1,05	0,28	Diterima
14.	1,02	0,33	Diterima
15.	1,06	0,27	Diterima
16.	0,94	-0,05	Diterima
17.	0,91	-0,75	Diterima
18.	0,99	0,30	Diterima
19.	1,11	0,43	Diterima
20.	1,05	0,27	Diterima

Hasil reliabilitas dilihat pada *Output Summary Statistics* dengan memperhatikan nilai *Pearson Reliability* dan *item reliability*. Hasil reliabilitas keseluruhan

butir soal instrumen *four tier diagnostic test* dapat dilihat pada Gambar 1 dan Gambar 2.

SUMMARY OF 25 MEASURED Person								
	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	6.0	20.0	-1.21	.59	1.00	-.15	1.02	-.09
SEM	.3	.0	.11	.01	.10	.30	.15	.25
P.SD	1.5	.0	.54	.05	.47	1.47	.73	1.23
S.SD	1.5	.0	.55	.05	.48	1.50	.74	1.26
MAX.	9.0	20.0	-.24	.70	2.04	2.56	3.04	2.07
MIN.	3.0	20.0	-2.36	.54	.44	-2.35	.34	-1.81

REAL RMSE	.65	TRUE SD	.00	SEPARATION	.00	Person RELIABILITY	.00	
MODEL RMSE	.59	TRUE SD	.00	SEPARATION	.00	Person RELIABILITY	.00	
S.E. OF Person MEAN = .11								

Person RAW SCORE-TO-MEASURE CORRELATION = 1.00 (approximate due to missing data)								
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .00 SEM = 1.48								
approximate due to missing data) STANDARDIZED (50 ITEM) RELIABILITY = .00								

Gambar 1. *Pearson Reliability*

SUMMARY OF 20 MEASURED Item								
	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	7.5	25.0	.00	.59	1.00	.03	1.02	.10
SEM	1.4	.0	.32	.05	.02	.09	.04	.12
P.SD	6.0	.0	1.40	.20	.08	.41	.19	.53
S.SD	6.2	.0	1.44	.20	.08	.42	.19	.54
MAX.	19.0	25.0	2.08	1.02	1.11	.43	1.45	.92
MIN.	1.0	25.0	-2.43	.42	.77	-1.11	.73	-1.12
REAL RMSE	.63	TRUE SD	1.25	SEPARATION	1.99	Item	RELIABILITY	.80
MODEL RMSE	.62	TRUE SD	1.26	SEPARATION	2.02	Item	RELIABILITY	.80
S.E. OF Item MEAN = .32								
Item RAW SCORE-TO-MEASURE CORRELATION = -.98 (approximate due to missing data)								
Global statistics: please see Table 44. UMEAN=.0000 USCALE=1.0000								

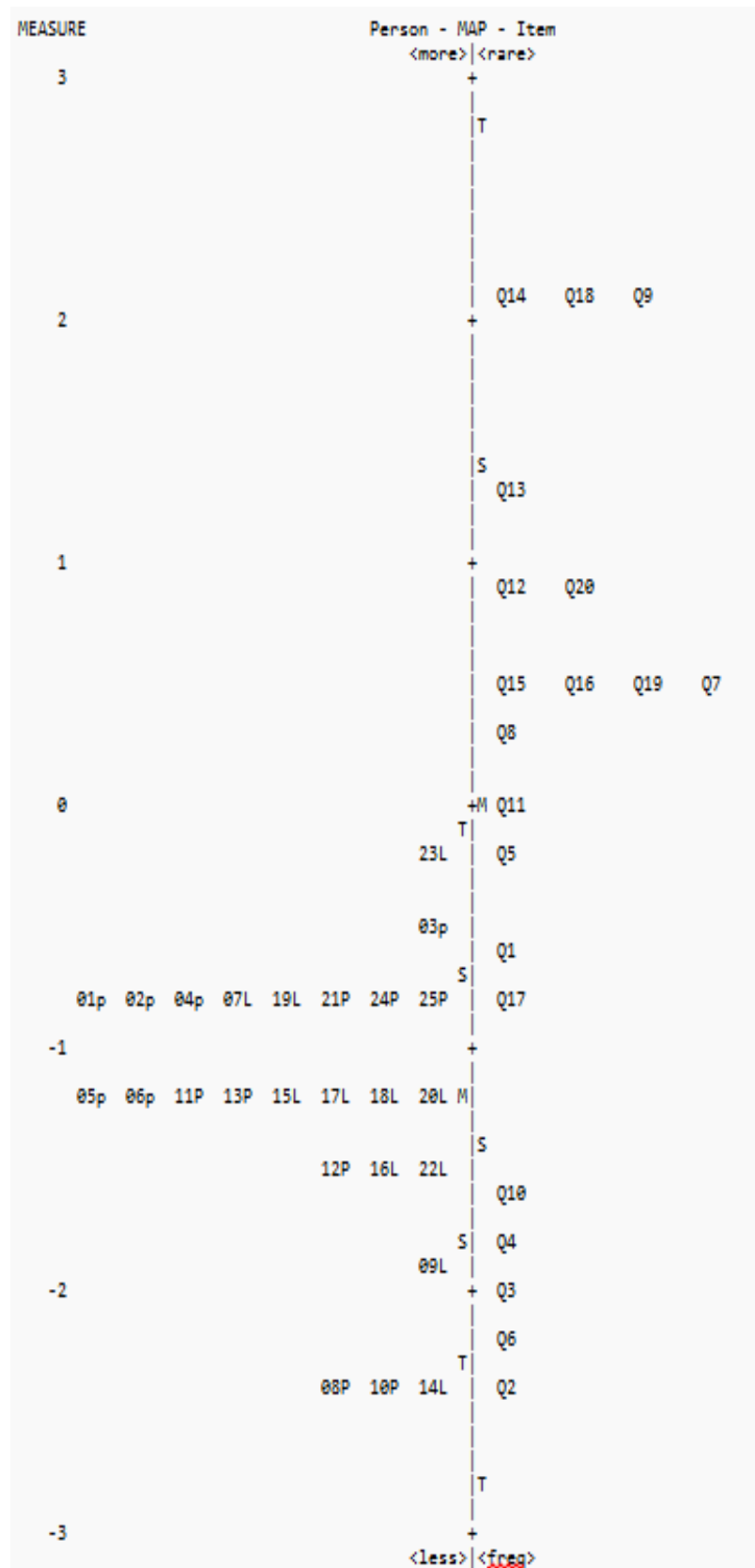
Gambar 2. Item Reliability

Gambar 1 menjelaskan hasil dari *Pearson Reliability* bernilai 0,00 sehingga masuk dalam kategori lemah dan nilai *alfa cronbach* 0,00 masuk dalam kategori buruk, sedangkan Gambar 2 *item reliability* 0,80 masuk dalam kategori bagus. Hal Ini menandakan bahwa lemahnya konsistensi peserta didik dalam menjawab butir-butir soal dan buruknya interaksi antara *pearson* dan *item* butir soal secara keseluruhan, namun kualitas dari butir-butir soal pada instrumen *four-tier diagnostic test* yang dikembangkan aspek reliabilitasnya bagus, sehingga butir-butir soal pada instrumen tersebut dapat berfungsi dengan baik dalam pengukuran dan reliabel. Seperti yang dikatakan oleh Sumintono & Widhiarso (2015), reliabel itu sendiri artinya suatu butir soal yang sudah dilakukan pengujian ke peserta didik memperoleh hasil yang konsisten.

Analisis ketiga adalah tingkat kesukaran soal yang dikembangkan. Pada Gambar 3 menunjukkan sebaran abilitas peserta didik dalam menjawab butir soal dan juga sebaran

tingkat kesukaran butir soal. Kode huruf L atau P yang diikuti nomor pada sebelah kiri menunjukkan nomor urut peserta didik. Pada Gambar 12 bagian kiri kode 23L menunjukkan nomor urut peserta didik yang mempunyai abilitas tertinggi, dengan nilai 0,0 *logit* dan pada kode 08P, 10P, 14L menunjukkan nomor urut peserta didik yang mempunyai abilitas yang terendah, dengan nilai -2 *logit*.

Pada kode Q yang diikuti nomor pada bagian kanan Gambar 4 menunjukkan nomor butir soal dan. Dilihat pada gambar bagian kanan paling atas kode Q14, Q18, Q9 menunjukkan butir soal yang memiliki tingkat kesukaran yang paling tinggi, masing-masing dapat dijawab oleh 1 peserta didik dengan kode 06P, 01P, dan 04P dan kode butir soal Q2 menunjukkan butir soal yang memiliki tingkat kesukaran terendah, dapat dijawab oleh 19 peserta didik dengan kode 02P, 03P, 04P, 05P, 07P, 08P, 10P, 11P, 12P, 15L, 16L, 17L, 18L, 19L, 21P, 22L, 23L, 24P, 25P.



Gambar 3. Variabel maps tingkat kesukaran

Item STATISTICS: MEASURE ORDER													
ENTRY	TOTAL	TOTAL	JMLE	MODEL	INFIT		OUTFIT		PTMEASUR-AL		EXACT MATCH		Item
NUMBER	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
9	1	25	2.08	1.02	.99	.30	.78	.14	.14	.09	96.0	96.0	Q9
14	1	25	2.08	1.02	1.02	.33	1.06	.42	.02	.09	96.0	96.0	Q14
18	1	25	2.08	1.02	.99	.30	.78	.14	.14	.09	96.0	96.0	Q18
13	2	25	1.33	.74	1.05	.28	1.45	.79	-.09	.12	92.0	92.0	Q13
12	3	25	.87	.62	1.05	.26	1.09	.34	.03	.14	88.0	87.9	Q12
20	3	25	.87	.62	1.05	.27	1.05	.27	.03	.14	88.0	87.9	Q20
7	4	25	.53	.55	1.04	.23	.98	.10	.11	.16	84.0	83.9	Q7
15	4	25	.53	.55	1.06	.27	1.16	.48	.02	.16	84.0	83.9	Q15
16	4	25	.53	.55	.94	-.05	.83	-.25	.29	.16	84.0	83.9	Q16
19	4	25	.53	.55	1.11	.43	1.28	.73	-.11	.16	84.0	83.9	Q19
8	5	25	.25	.51	1.09	.39	1.32	.92	-.06	.18	80.0	79.9	Q8
11	6	25	.01	.48	1.04	.25	1.19	.68	.07	.19	76.0	75.9	Q11
5	7	25	-.21	.46	.94	-.23	.95	-.10	.29	.21	72.0	71.9	Q5
1	9	25	-.60	.43	.99	-.04	1.04	.29	.22	.23	72.0	64.8	Q1
17	10	25	-.77	.42	.91	-.75	.87	-.84	.41	.24	60.0	61.5	Q17
10	15	25	-1.64	.42	.95	-.36	.93	-.44	.36	.26	68.0	64.7	Q10
4	16	25	-1.82	.43	.91	-.52	.89	-.64	.42	.26	72.0	67.3	Q4
3	17	25	-2.01	.44	1.04	.27	1.01	.13	.20	.25	64.0	70.1	Q3
6	18	25	-2.21	.46	.77	-1.11	.73	-1.12	.66	.25	84.0	72.9	Q6
2	19	25	-2.43	.48	1.02	.15	.98	.03	.23	.24	76.0	76.0	Q2
MEAN	7.5	25.0	.00	.59	1.00	.03	1.02	.10			80.8	79.8	
P.S.D	6.0	.0	1.40	.20	.08	.41	.19	.53			10.3	10.8	

Gambar 4. Nilai *measure* setiap butir soal

Gambar 4 merupakan data tingkat kesukaran butir soal *four tier diagnostic test* yang dilihat dari nilai *measure* yang dapat dilihat pada *Output Tabel 13. Item Measure* yang diurutkan tingkat kesukaran butir soal dari tingkat kesukaran yang tertinggi ke yang terendah yang ditunjukkan pada kolom *measure* dalam satuan *logit* (Sumintono & Widhiarso, 2015:69). Pada Gambar 4 di atas butir soal yang

memiliki tingkat kesukaran tertinggi dimiliki oleh Q9, Q14, dan Q18 (*measure* = +2,08 *logit*) dan butir soal yang memiliki tingkat kesukaran terendah dimiliki oleh Q2 (*measure* = -2,43 *logit*). Berdasarkan Gambar 3 dan 4, tingkat kesukaran seluruh item butir soal pada instrumen *four tier diagnostic test* dijabarkan dalam bentuk tabel dapat dilihat pada Tabel 8.

Tabel 8. Hasil keseluruhan tingkat kesukaran butir soal

No. Soal	Tingkat Kesukaran Soal	Keterangan
1.	-0,60	Mudah
2.	-2,43	Sangat Mudah
3.	-2,01	Sangat Mudah
4.	-1,82	Sangat Mudah
5.	-0,21	Mudah
6.	-2,21	Sangat Mudah
7.	0,53	Sukar
8.	0,25	Sukar
9.	2,08	Sangat Sukar
10.	-1,64	Sangat Mudah
11.	0,01	Sukar
12.	0,87	Sukar
13.	1,33	Sangat Sukar
14.	2,08	Sangat Sukar
15.	0,53	Sukar
16.	0,53	Sukar
17.	-0,77	Mudah
18.	2,08	Sangat Sukar
19.	0,53	Sukar
20.	0,87	Sukar

Dilihat dari Tabel 8 menunjukkan instrumen *four tier diagnostic test* yang dikembangkan mempunyai tingkat kesukaran yang beragam, dimulai dari butir soal sangat mudah, mudah, sukar, dan sangat sukar. Sehingga penyebaran tingkatan kesukaran butir soal yang dikembangkan dikatakan Cukup merata. Ini sesuai dengan pendapat Erfan et al. (2020:12), soal yang tidak terlalu sukar dan juga tidak terlalu mudah dikatakan sebagai soal yang baik kualitasnya. Seperti pendapatnya Rahmasari & Ismiyati (2016:321), butir soal yang memiliki tingkat kesukaran yang tinggi membuat peserta didik putus asa dan tidak

semangat untuk mencoba memecahkan permasalahan dalam soal dan sebaliknya jika butir soal yang memiliki tingkat kesukaran yang sangat mudah.

Daya Beda

Hasil nilai daya beda pada seluruh butir soal dapat dilihat dari *Item (Column): Fit Order* dengan memperhatikan nilai PTMEASURE-AL COOR pada seluruh butir soal *four-tier diagnostic test*. Hasilnya dapat dilihat pada Gambar 5, sehingga daya beda pada seluruh butir soal dapat dilihat pada Tabel 9.

Item STATISTICS: MISFIT ORDER													
ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	OUTFIT MNSQ	PTMEASURE-AL CORR.	EXACT EXP.	EXACT OBS%	MATCH EXP%	Item		
13	2	25	1.33	.74	1.05	.28	1.45	.79	A-.09	.12	92.0	92.0	Q13
8	5	25	.25	.51	1.09	.39	1.32	.92	B-.06	.18	80.0	79.9	Q8
19	4	25	.53	.55	1.11	.43	1.28	.73	C-.11	.16	84.0	83.9	Q19
11	6	25	.01	.48	1.04	.25	1.19	.68	D .07	.19	76.0	75.9	Q11
15	4	25	.53	.55	1.06	.27	1.16	.48	E .02	.16	84.0	83.9	Q15
12	3	25	.87	.62	1.05	.26	1.09	.34	F .03	.14	88.0	87.9	Q12
14	1	25	2.08	1.02	1.02	.33	1.06	.42	G .02	.09	96.0	96.0	Q14
20	3	25	.87	.62	1.05	.27	1.05	.27	H .03	.14	88.0	87.9	Q20
1	9	25	-.60	.43	.99	-.04	1.04	.29	I .22	.23	72.0	64.8	Q1
3	17	25	-2.01	.44	1.04	.27	1.01	.13	J .20	.25	64.0	70.1	Q3
7	4	25	.53	.55	1.04	.23	.98	.10	j .11	.16	84.0	83.9	Q7
2	19	25	-2.43	.48	1.02	.15	.98	.03	i .23	.24	76.0	76.0	Q2
9	1	25	2.08	1.02	.99	.30	.78	.14	h .14	.09	96.0	96.0	Q9
18	1	25	2.08	1.02	.99	.30	.78	.14	g .14	.09	96.0	96.0	Q18
5	7	25	-.21	.46	.94	-.23	.95	-.10	f .29	.21	72.0	71.9	Q5
10	15	25	-1.64	.42	.95	-.36	.93	-.44	e .36	.26	68.0	64.7	Q10
16	4	25	.53	.55	.94	-.05	.83	-.25	d .29	.16	84.0	83.9	Q16
4	16	25	-1.82	.43	.91	-.52	.89	-.64	c .42	.26	72.0	67.3	Q4
17	10	25	-.77	.42	.91	-.75	.87	-.84	b .41	.24	60.0	61.5	Q17
6	18	25	-2.21	.46	.77	-1.11	.73	-1.12	a .66	.25	84.0	72.9	Q6

MEAN	7.5	25.0	.00	.59	1.00	.03	1.02	.10			80.8	79.8	
P.SD	6.0	.0	1.40	.20	.08	.41	.19	.53			10.3	10.8	

Gambar 5. Nilai PTMEASURE-AL COOR

Tabel 9. Hasil PTMEASURE-AL COOR

No. Soal	PTMEASURE-AL COOR	Keterangan
1.	0,22	Kurang baik
2.	0,23	Kurang baik
3.	0,20	Kurang baik
4.	0,42	Sangat baik
5.	0,29	Kurang baik
6.	0,66	Sangat baik
7.	0,11	Kurang baik
8.	-0,06	Tidak baik
9.	0,14	Tidak baik
10.	0,36	Baik
11.	0,07	Tidak baik
12.	0,03	Tidak baik
13.	-0,09	Tidak baik
14.	0,02	Tidak baik
15.	0,02	Tidak baik

No. Soal	PTMEASURE-AL COOR	Keterangan
16.	0,29	Kurang Baik
17.	0,41	Sangat Baik
18.	0,14	Tidak baik
19.	-0,11	Tidak baik
20.	0,03	Tidak baik

Pada Tabel 9 menunjukkan hasil bahwa terdapat 10 butir soal (Q8, Q9, Q11, Q12, Q13, Q14, Q15, Q18, Q19, Q20) yang mempunyai daya pembeda yang masuk dalam kategori tidak baik, 3 butir soal (Q8, Q13, Q19) diantaranya bernilai *negative*. Selain itu terdapat 6 butir soal (Q1, Q2, Q3, Q5, Q7, Q16) yang masuk dalam kategori kurang baik, 1 butir soal (Q10) dengan kategori baik, dan 3 butir soal (Q4, Q6, Q17) dengan kategori sangat baik. Soal yang mendapatkan nilai PTMEASURE-AL COOR *negative*, dibuang atau tidak digunakan pada uji coba skala besar, sehingga soal yang nantinya digunakan hanya 17 butir soal.

Pembahasan

Dilihat dari hasil validitas setiap butir soal pada instrumen *four-tier diagnostic test* menunjukkan hasil keseluruhannya memiliki nilai MNSQ bernilai 1,05 sampai 0,77 sedangkan nilai ZSTD memiliki nilai 0,28 *logit* sampai dengan 1,11 *logit* sehingga secara keseluruhan seluruh item butir soal sesuai dengan kriteria dimana nilai *Outfit Mean Square (MNSQ)* lebih besar dari 0,5 *logit* dan lebih kecil dari 1,5 *logit*. Untuk itu setiap butir soal nomor 1 sampai 20 dapat digunakan atau dapat dikatakan pola respon peserta didik untuk keseluruhannya dinilai fit hal ini berarti keseluruhan butir soal *four-tier diagnostic test* berfungsi secara normal atau valid. Hal ini sejalan dengan pendapat Rohmah et al. (2022), mengatakan butir soal yang dapat berfungsi dengan baik atau tidak dalam pengukuran dilihat dari *item fit* pada setiap butir soal. Selain itu juga dijelaskan oleh Az-Zahra et al. (2024:13), jika nilai yang didapatkan semakin mendekati angka, maka item tersebut sesuai dengan kategori dan dikatakan item tersebut fit.

Reliabilitas adalah indeks yang menunjukkan sejauh mana instrumen *four tier diagnostic test* tersebut dapat dipercaya (Tarigan et al. 2022:95). Pada hasil uji reliabilitas yang dianalisis dengan menggunakan minitab versi 5.2.3 dilihat dari *Output Table 3.1 Summary Statistic* dengan melihat nilai *person reliability* dan *item reliability*. Hasil dari *person reliability* 0,00

masuk dalam kategori lemah dan nilai *alfa cronbach* 0,00 masuk dalam kategori buruk, sedangkan *item reliability* 0,80 masuk dalam kategori bagus. Ini menunjukkan bahwa lemahnya konsistensi peserta didik dalam menjawab butir-butir soal dan buruknya interaksi antara *person* dan item butir soal secara keseluruhan, namun kualitas dari butir-butir soal pada instrumen *four-tier diagnostic test* yang dikembangkan aspek reliabilitasnya bagus, sehingga butir-butir soal pada instrumen tersebut dapat berfungsi dengan baik dalam pengukuran dan reliabel. Seperti yang dikatakan oleh Sumintono & Widhiarso (2015:10), reliabel itu sendiri artinya suatu butir soal yang sudah dilakukan pengujian ke peserta didik memperoleh hasil yang konsisten.

Dari hasil analisis pada butir soal peserta didik kode 23L mempunyai abilitas tertinggi, dengan nilai 0,0 *logit* dan pada kode 08P, 10P, 14L menunjukkan abilitas yang terendah, dengan nilai -2 *logit*. Peserta didik dengan kode Q14, Q18, Q9 (*measure* = +2,08 *logit*) menunjukkan butir soal yang memiliki tingkat kesukaran yang paling tinggi, masing-masing dapat dijawab oleh 1 peserta didik dengan kode 06P, 01P, dan 04P dan kode butir soal Q2 (*measure* = -2,43 *logit*) menunjukkan butir soal yang memiliki tingkat kesukaran terendah, dapat dijawab oleh 19 peserta didik dengan kode 02P, 03P, 04P, 05P, 07P, 08P, 10P, 11P, 12P, 15L, 16L, 17L, 18L, 19L, 21P, 22L, 23L, 24P, 25P. Dari Tabel 4.6 menunjukkan instrumen yang dikembangkan mempunyai tingkat kesukaran yang beragam, dimulai dari butir soal sangat mudah, mudah, sukar, dan sangat sukar. Sehingga penyebaran tingkatan kesukaran butir soal yang dikembangkan dikatakan Cukup merata. Ini sesuai dengan pendapat Erfan et al. (2020:12), soal yang tidak terlalu sukar dan juga tidak terlalu mudah dikatakan sebagai soal yang baik kualitasnya.

Hasil dari uji daya beda terdapat 10 butir soal (Q8, Q9, Q11, Q12, Q13, Q14, Q15, Q18, Q19, Q20) yang mempunyai daya pembeda yang masuk dalam kategori jelek, 3 butir soal (Q8, Q13, Q19) diantaranya bernilai *negative*. Selain itu terdapat 6 butir soal (Q1, Q2, Q3, Q5,

Q7, Q16) yang masuk dalam kategori kurang baik, 1 butir soal (Q10) dengan kategori baik, dan 3 butir soal (Q4, Q6, Q17) dengan kategori sangat baik. Soal yang mendapatkan nilai PTMEASURE-AL COOR *negative*, dibuang atau tidak digunakan pada uji coba skala besar. Sehingga soal yang nantinya digunakan hanya 17 butir soal. Hal ini mengacu pada pendapat Smiley (2015:20), mengatakan bahwa nilai negatif menunjukkan butir soal yang jelek atau cacat karena peserta didik yang memiliki abilitas yang rendah lebih sering benar dalam menjawab butir soal dibandingkan dengan peserta didik yang memiliki abilitas yang tinggi, selain itu menurutnya nilai-nilai negatif yang dihasilkan lebih bermasalah daripada nilai nol, sehingga soal tersebut dapat direvisi atau mungkin dihapus. Pendapat ini sejalan dengan pendapat Jumini et al. (2023:763), mengatakan bahwa daya beda pada butir soal digunakan untuk dapat membedakan peserta didik yang memiliki kemampuan tinggi dan kemampuan yang rendah dalam menjawab butir soal.

SIMPULAN

Hasil penelitian dan pengembangan ini dapat disimpulkan bahwa Instrumen *four-tier diagnostic test* pada materi ikatan kimia menunjukkan validitas setiap butir soal telah sesuai dengan kriteria. Reliabilitas instrumen berdasarkan *person reliability* adalah 0,00 *logit* (kategori lemah) dan nilai alfa Cronbach sebesar 0,00 *logit* (kategori buruk). Sementara itu, *item reliability* mencapai 0,80 *logit* (kategori bagus). Tingkat kesukaran soal bervariasi, mulai dari sangat mudah, mudah, sukar, hingga sangat sukar. Daya pembeda menunjukkan 10 butir soal berkategori jelek, dengan 3 di antaranya bernilai negatif. Sebanyak 6 butir soal masuk kategori kurang baik, 1 soal kategori baik, dan 3 soal kategori sangat baik.

DAFTAR PUSTAKA

- Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, 40(4), 955–959. <https://doi.org/10.1177/001316448004000419>
- Atkins, P. (2013). *What is Chemistry?* (1st ed.). Oxford University Press. <https://en.id1lib.org/book/2616269/8ff094>
- Az-Zahra, N., Suryana, D., & Nadhirah, N. A. (2024). Analisis validitas konstruk instrumen kematangan karier menggunakan winstep. *Jurnal Bimbingan Dan Konseling Ar-Rahman*, 10(1), 8–20. <https://doi.org/10.31602/jbkr.v10i1.13745>
- Coll, R. K., & Taylor, N. (2002). Mental models in chemistry: senior chemistry students' mental models of chemical bonding. *Chem. Educ. Res. Pract.*, 3(2), 175–184. <https://doi.org/10.1039/B2RP90014A>
- Derkach, T. M. (2021). The origin of misconceptions in inorganic chemistry and their correction by computer modelling. *Journal of Physics: Conference Series*, 1840(1). <https://doi.org/10.1088/1742-6596/1840/1/012012>
- Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). Analisis kualitas soal kemampuan membedakan rangkaian seri dan paralel melalui teori tes klasik dan model rasch. *Indonesian Journal of Educational Research and Review*, 3(1), 11–19.
- Frahiwi, N. J., Kaniawati, I., Suhendi, E., Suyana, I., & Samsudin, A. (2017). The transformation of two-tier into four-tier test on newton's laws concepts. *AIP Conference Proceedings*, 1848(1), 3–8. <https://doi.org/10.1063/1.4983967>
- Habiddin, & Page, E. M. (2019). Development and validation of a four-tier diagnostic instrument for chemical kinetics (FTDICK). *Indones. J. Chem*, 19(3), 720–736. <https://doi.org/10.22146/ijc.39218>
- Hadinugrahaningsih, T., Andina, R. E., Munggaran, L. R., & Rahmawati, Y. (2020). Analysis of students' alternative conceptions about electrolyte and non-electrolyte solutions using a two-tier diagnostic test for chemistry teaching improvement. *Universal Journal of Educational Research*, 8(5), 1926–1934. <https://doi.org/10.13189/ujer.2020.080529>
- Harahap, I. P. P., & Novita, D. (2020). Validitas dan Reliabilitas Instrumen Tes Diagnostik Four-Tier Multiple Choice (4TMC) Pada Konsep Laju Reaksi. *Unesa Journal of Chemical Education*, 9(2), 222–227.
- Härmälä-Braskén, A. S., Hemmi, K., & Kurtén, B. (2020). Misconceptions in Chemistry Among Finnish Prospective Primary School Teachers—a Long-Term Study. *International Journal of Science Education*, 42(9), 1447–1464. <https://doi.org/10.1080/09500693.2020.17>

- 65046
- Jumini, S., Madnasri, S., Cahyono, E., & Parmin, P. (2023). Analisis Kualitas Butir Soal Pengukuran Literasi Sains Melalui Teori Tes Klasik Dan Rasch Model. *Prosiding Seminar Nasional Pascasarjana Universitas Negeri Semarang*, 758–765.
- Murniningsih, Muna, K., & Irawati, R. K. (2020). Analysis of Misconceptions by Four Tier Tests in Electrochemistry, Case Study on Students of the Chemistry Education Study Program UIN Antasari Banjarmasin. *Journal of Physics: Conference Series*, 1440, 1–6. <https://doi.org/10.1088/1742-6596/1440/1/012008>
- Özmen, H. (2004). Some Student Misconceptions in Chemistry: A Literature Review of Chemical Bonding. *Journal of Science Education and Technology*, 13(2), 147–159. <https://doi.org/10.1023/B:JOST.0000031255.92943.6d>
- Pratama, F. I., & Rohaeti, E. (2024). How Does “Chemistry Challenge” E-book Affect the Chemical Literacy Profile? A Study to Test Learning Media Effectiveness. *Jurnal Penelitian Pendidikan IPA*, 10(5), 2253–2260. <https://doi.org/10.29303/jppipa.v10i5.7018>
- Purniasari, L., Masykuri, M., & Ariani, S. R. D. (2021). Analisis Butir Soal Ujian Sekolah Mata Pelajaran Kimia SMAN 1 Kutowinangun Tahun Pelajaran 2019/2020 Menggunakan Model Iteman dan Rasch. *Jurnal Pendidikan Kimia*, 10(2), 205–214. <https://jurnal.uns.ac.id/jpkim>
- Rahayu, I. (2009). *Praktis Belajar Kimia untuk Kelas X Sekolah Menengah Atas/Madrasah Aliyah* (F. Dzalfa (ed.)). Jakarta: Visindo Media Persada.
- Rahmasari, D., & Ismiyati. (2016). Analisis Butir Soal Mata Pelajaran Pengantar Administrasi Perkantoran. *Economic Education Analysis Journal*, 5(1), 317–330.
- Rahmiati, Irawati, R. K., & Astutik, T. P. (2022). Pemahaman Konsep Pada Materi Reaksi Redoks dengan Four-Tier Test. *AL KAWNU: SCIENCE AND LOCAL WISDOM JOURNAL*, 01(02), 13–17. <https://doi.org/10.18592/alkawnu.v1i1.6343>
- 43
- Rohmah, F. N., Susilaningsih, E., Haryani, S., & Kasmui. (2022). Desain Asesmen Kompetensi Minimum Literasi Membaca Bermuatan High Order Thinking Skills untuk Menganalisis Kompetensi Minimum Siswa Materi Asam-Basa. *Chemistry in Education*, 11(2), 117–125.
- Sesen, B. A. (2010). Internet as a Source of Misconception: “Radiation and Radioactivity.” *TOJET: The Turkish Online Journal of Educational Technology*, 9(4), 94–100. <https://eric.ed.gov/?id=EJ908075>
- Setiawan, C. E., & Ilahi, P. R. (2022). Identification of Misconceptions in Chemical Bonding Materials Using Three Tier Diagnostic Test. *Journal of Natural Science and Integration*, 5(1), 77. <https://doi.org/10.24014/jnsi.v5i1.16860>
- Smiley, J. (2015). Classical Test Theory or Rasch: A Personal Account from a Novice User. *Shiken*, 19(1), 16–29. <https://www.researchgate.net/publication/288858190>
- Stefani, C., & Tsaparlis, G. (2009). Students’ Levels of Explanations, Models, and Misconceptions in Basic Quantum Chemistry: A Phenomenographic Study. *Journal of Research in Science Teaching*, 46(5), 520–536. <https://doi.org/10.1002/tea.20279>
- Sugiyono. (2010). *Metode Penelitian Pendidikan: Pendekatan Kuantitatif, Kualitatif, dan R&D*. Bandung: Alfabeta.
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. Cimahi: Trim Komunikata.
- Taber, K. S. (2009). Challenging Misconceptions in the Chemistry Classroom: Resources to Support Teachers. *Educació Química EduQ*, 4, 13–20. <https://doi.org/10.2436/20.2003.02.27>
- Tarigan, E. F., Nilmarito, S., Islamiyah, K., Darmana, A., & Suyanti, R. D. (2022). Analisis Instrumen Tes Menggunakan Rasch Model dan Software SPSS 22.0. *Jurnal Inovasi Pendidikan Kimia*, 16(2), 92–96.
- Tümay, H. (2016a). Emergence, Learning Difficulties, and Misconceptions in Chemistry Undergraduate Students’ Conceptualizations of Acid Strength. *Science and Education*, 25(1–2), 21–46.

<https://doi.org/10.1007/s11191-015-9799-x>

- Tümay, H. (2016b). Reconsidering Learning Difficulties and Misconceptions in Chemistry: Emergence in Chemistry and its Implications For Chemical Education. *Chemistry Education Research and Practice*, 17(2), 229–245. <https://doi.org/10.1039/c6rp00008h>
- Üce, M., & Ceyhan, İ. (2019). Misconception in Chemistry Education and Practices to Eliminate Them: Literature Analysis. *Journal of Education and Training Studies*, 7(3), 202–208.

<https://doi.org/10.11114/jets.v7i3.3990>

PROFIL SINGKAT

Munawaroh merupakan mahasiswa S2 Pendidikan Kimia Universitas Negeri Yogyakarta. Dapat dihubungi melalui alamat email: munawaroh2802@gmail.com

Das Salirawati merupakan dosen Pendidikan Kimia Universitas Negeri Yogyakarta dengan bidang keahlian Penilaian dan Evaluasi Pendidikan. Dapat dihubungi melalui alamat email: das_salirawati@uny.ac.id