E-ISSN: 2798-0472 DOI: 10.52562/biochephy.v5i1.1466

BIOCHEPHY: Journal of Science Education

http://journal.moripublishing.com/index.php/biochephy

DEVELOPMENT OF A FIVE TIER DIAGNOSTIC TEST OF MISCONCEPTIONS ON CHEMICAL EQUILIBRIUM MATERIAL

Anwar Saputra, Hesty Parbuntari*
Universitas Negeri Padang, Indonesia
*Corresponding author: hesty5193@fmipa.unp.ac.id

Abstract: Misconceptions in learning chemistry have the potential to have a negative impact because of the close interrelationship of chemical materials. This study aims to develop an effective five-level diagnostic test to identify students' misconceptions in the topic of chemical equilibrium. Through R&D development and Rasch model. Data from validation by experts (3 chemistry lecturers and 2 high school teachers) were processed using Minifac software. Furthermore, the raw data from the pilot test on 32 students were analyzed with Ministep software. The results of the analysis showed that the research instrument had good content validity, sufficient to very good reliability (0.70-0.86 The analysis of the difficulty index shows that there are three categories of questions at the first, third and fifth levels. Differentiated power also showed a similar pattern, with three categories at the first, third and fifth levels.

Keywords: Five-Tier Test Instrument, Misconceptions, Chemical Equilibrium, Rasch Model

***Abstrak**: Miskonsepsi dalam pembelajaran kimia berpotensi menimbulkan dampak negative karena keterkaitan materi kimia yang erat. Penelitian ini bertujuan mengembangkan tes diagnostik lima tingkatan yang efektif untuk mengidentifikasi miskonsepsi peserta didik dalam topic kesetimbangan kimia. Melalui pengembangan R&D dan model Rasch. Data hasil validasi oleh para ahli (3 dosen kimia dan 2 guru SMA) diproses menggunakan perangkat lunak Minifac. Selanjutnya, data mentah hasil uji coba pada 32 peserta didik dianalisis dengan software Ministep. Hasil analisis menunjukkan bahwa instrument penelitian memiliki validitas isi yang baik, reliabilitas yang cukup hingga sangat baik (0.70-0.86), analisis indeks kesukaran menunjukkan bahwa soal-soal pada tingkat pertama, ketiga dan kelima terdapat tiga kategori. Daya beda soal juga menunjukkan pola yang serupa, dengan tiga kategori pada tingkat pertama, ketiga dan kelima.*

***Kata Kunci:** Instrument Test Five-Tier ,Miskonsepsi, Kesetimbangan Kimia Model rasch*

INTRODUCTION

Chemistry is a branch of science that focuses on an in-depth understanding of matter, from its atomic properties and structure to the changes and reactions that occur, as well as the energy that accompanies each change. The abstract and complex nature of some chemical concepts causes students to have difficulty in understanding the material because the concept cannot be observed directly and requires a gradual and thorough understanding(Agustin et al., 2022; Zulfadli & Munawwarah, 2016).

Learners face difficulties that result in them having inappropriate understandings. When learners have a different and inaccurate understanding of a chemical concept, this is called misconception. These misconceptions are often a barrier to deeper understanding (Agustin et al., 2022). Misconceptions refer to learners' understanding of a concept that is not in accordance with scientific concepts recognized by experts. In other words, misconceptions can be understood

as a mismatch between learners' understanding and the understanding that has been generally accepted by experts (Permatasari et al., 2022).

Chemistry learning is often hampered by misconceptions that produce unintended consequences. This is due to the interconnected nature of chemical materials. One topic that often causes difficulties and has a high risk of causing misconceptions is chemical equilibrium, which is considered a complex concept in the teaching and learning process (Monita, Ade & Bambang, 2016).

There are various ways to identify misconceptions in students, one of the efficient strategies to identify students' misconceptions is through the application of diagnostic tests, this is because diagnostic tests have proven to be effective in uncovering misconceptions that occur during the learning process. The right diagnostic test can provide clear and accurate information about students' misconceptions, based on the mistakes they make (Agustin et al., 2022; Monita, Ade & Bambang, 2016).

One form of diagnostic test is a multiple-choice test, which is often known as a multiple-choice diagnostic test. Misconceptions cannot be detected using only a single-level multiple-choice test, the single-level multiple-choice diagnostic test has limitations in identifying misconceptions due to chance or guessing factors, making it difficult to understand the reasons behind their understanding. For that reason, further development on multiple-choice tests is needed. The latest and most recent instrument development is a five-tier test instrument that has been carried out by on the concept of sound waves (Lailiyah & Ermawati, 2020), on reaction rate material (Nisa & Sudrajat, 2023) and on harmonic vibration material(Putri & Ermawati, 2021).

The five tier diagnostic test is a form of development of the four tier test instrument, where in this four tier students are still likely to be able to guess the answer, so a five tier test instrument was developed to reduce guessing in the answer (Fajriyyah & Ermawati, 2020). The five tier diagnostic test consists of five tiers, the first tier includes objective questions and answer choices, the second tier is related to the confidence of the answer, the third tier explains the reason behind the answer in the first tier, the fourth tier assesses the level of confidence of the reason, and the fifth tier consists of additional open-ended questions in the form of drawing tests, inferences or other types of tests tailored to the needs of each item (Lailiyah & Ermawati, 2020; Nisa & Sudrajat, 2023; Putri & Ermawati, 2021; Setiawan & Faoziyah, 2020).

METHOD

Explaining research chronological, including research design, research procedure, how to test the data The research was conducted at SMA Pembangunan Universitas Negeri Padang on November-December 2024, the research used in this study was Research and Development (research and development) with the Rasch model. This Rasch model can increase the accuracy of statistical results in the analysis carried out (Sumintono & Widhiarso, Aplikasi Model Rasch Untuk Penelitian ilmu-ilmu Sosial, 2014). The research procedure carried out with the Rasch model uses 10 specified steps (Liu, 2012) (1) State the objectives and the intended population (2) Determine the constructs to be measured. At this stage, it is done by analyzing the Learning Outcomes (CP), Learning Objectives (TP), and Flow of Learning Objectives (ATP) on the topic of Chemical Equilibrium material. (3) Creating question indicators from the specified construct. At this stage, it is done by compiling question indicators in accordance with the analysis carried out. After making the question indicators, then design the question items, the answer key and the assessment rubric. (4) Conduct a trial test. At this stage the instrument is tested in accordance with the specified subject. (5) Perform analysis using the Rasch model. (6) Reviewing item fit statistics and revising items if necessary. (7) Reviewing the wright map and adding/deleting items, if necessary. (8) Repeating the procedure of steps 4-7 until all items fit the Rasch model. (9) Determine the validity, reliability, difficulty index and differential power of the items. (10) Develop documentation for the test instrument.The data analysis carried out in the Rasch model consists of;

1. Test validity, In valid Rasch modeling, content assessment and test instruments can be evaluated through item fit analysis. This allows us to evaluate the extent to which each item fits the model used (Sumintono & Widhiarso, 2015). The validity carried out is by expert assessment which consists of chemistry lecturers and chemistry teachers, along with data obtained in the field based on the responses of students to the answers to each question. The standard used to see the level of item fit can be seen in the outfit (outler sensitive fit) which means measuring the sensitivity of the response pattern or vice versa.

Table 1. Criteria for Question Item Quality

Value Criteria	Quality Rating scale
Outfit means square (MNSQ)	0.5<MNSQ<1.5
Outfit Z-standard (ZSTD)	-2.0<ZSTD<+2.0
Point Measure Correlation (Pt mean Corr)	0.4< Pt Measure Corr< 0.85

2. Reliability Reliability on an instrument shows how reliable the instrument is in collecting data. There are several standards for assessing these reliability items, namely (Sumintono & Widhiarso, 2014).

Table 2 Criteria for item quality Reliability

Criteria	Rating scale quality
Weak	<0.67
Simply	0.67-0.80
Good	0.81-0.90
Very good	0.91-0.94
Special	>0.94

3. Item difficulty index In the Rasch model, this analysis is done through the item measure analysis menu feature. The criteria for the logit value on the item measure is that the highest logit will indicate a question item with high difficulty. While the lowest logit value will indicate difficulty (Sumintono & Widhiarso, 2015).

Table 3 Quality of Question Difficulty Index

Rating Scale Quality	Classification of Question Items
< -1SD	Very easy
0.0 *logit* - 1SD	Easy
0.0 *logit* + 1SD	Difficult
> 1SD	Very difficult

4. In the Rasch model, this analysis is carried out using the output table menu with the summary statistic option. The difference power of the summary statistic can identify the respondent group as well as the difference power group of the question. The differential power of the question can be seen from the separation value. The higher the separation value on the question, the better the differentiating power of the instrument. The equation used to be able to see the grouping more thoroughly, can use strata separation.

$$H = \frac{[(4\,x\,\text{SEPARATION})+1]}{3}$$ (Sumintono & Widhiarso, 2014)

The differential power is seen from the separation value which is calculated again using the H formula. The H value obtained in the form of an integer will have four categories with a distribution of very difficult, difficult, moderate and easy differential power. When the H value is

0, the instrument has no differentiation, so the quality of an instrument is not good (Sumintono & Widhiarso, 2014).

RESULT AND DISCUSSION

This study involved 5 experts with 3 Chemistry lecturers at Padang State University and 2 teachers at SMA Pembangunan Universitas Negeri Padang in product assessment and small-scale trials at Padang State University Development High School, in phase F. From the research that has been carried out, an assessment of the five-tier diagnostic test instrument for misconceptions in students of chemical equilibrium material is obtained. The research procedure was carried out using the Rasch model, by developing test instruments through predetermined steps (Liu, 2012) The results of this study obtained data that.

a. Validity

1) Logical validity

The five-tier diagnostic test instrument was validated by 5 experts consisting of 3 chemistry lecturers at Padang State University and 2 chemistry teachers at SMA Pembangunan Universitas Negeri Padang using a validation assessment sheet that had been prepared, each item had an assessment aspect from the validator. From the results of the validation of the five-tier diagnostic test instrument by five experts, it was found that the entire instrument was valid as a whole. The validation results can be seen in table 5 below.

Table 4 Expert Validation Results

Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu Item
64	65	.98	.98	.49	1.02	.96	.2	.42	.0	1.05	.18	.11	10 S10
64	65	.98	.98	.49	1.02	.99	.3	.53	.1	1.03	.16	.11	7 S7
63	65	.97	.97	-.24	.73	.95	.1	.54	-.1	1.06	.23	.16	15 S15
61	65	.94	.93	-1.01	.54	.95	.0	.67	-.3	1.06	.29	.22	1 S1
63	65	.97	.97	-.24	.73	1.01	.2	.72	.0	1.01	.18	.16	2 S2
64	65	.98	.98	.49	1.02	1.01	.3	.72	.2	1.00	.13	.11	11 S11
64	65	.98	.98	.49	1.02	1.01	.3	.72	.2	1.00	.13	.11	12 S12
64	65	.98	.98	.49	1.02	1.01	.3	.72	.2	1.00	.13	.11	13 S13
61	65	.94	.93	-1.01	.54	1.03	.2	.94	.1	.98	.22	.22	3 S3
64	65	.98	.98	.49	1.02	1.03	.3	1.09	.5	.97	.09	.11	6 S6
62	65	.95	.95	-.68	.61	1.07	.2	1.14	.4	.94	.15	.19	5 S5
63	65	.97	.97	-.24	.73	1.03	.2	1.35	.6	.96	.12	.16	4 S4
64	65	.98	.98	.49	1.02	1.06	.3	2.58	1.2	.90	.00	.11	14 S14
65	65	1.00	1.00	(1.71	1.84)	Maximum					.00	.00	8 S8
65	65	1.00	1.00	(1.71	1.84)	Maximum					.00	.00	9 S9
63.4	65.0	.98	.97	.23	.98	1.01	.3	.93	.3		.13		Mean (Count: 15)
1.2	.0	.02	.02	.79	.38	.04	.1	.54	.4		.08		S.D. (Population)
1.2	.0	.02	.02	.82	.40	.04	.1	.56	.4		.09		S.D. (Sample)

With extremes, Model, Populn: RMSE 1.05 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
With extremes, Model, Sample: RMSE 1.05 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Without extremes, Model, Populn: RMSE .87 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Without extremes, Model, Sample: RMSE .87 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
With extremes, Model, Fixed (all same) chi-squared: 10.4 d.f.: 14 significance (probability): .73

An instrument can be said to be valid if one of the following three criteria Outfit means square (MNSQ) 0.5 < MNSQ < 1.5, Outfit Z-standard (ZSTD) -2.0 < ZSTD < + 2.0, Point Measure Correlation (Pt mean Corr) 0.4 < Pt Measure Corr < 0.85. (Maulana et al., 2023; Sumintono & Widhiarso, 2014) From the results of the validation of the five-tier diagnostic test instrument by five experts, it was found that the entire instrument was valid overall. The validation results can be seen in table 4. In the MNSQ outfit criteria, it was found that there were several questions that did not enter the criteria where the criteria were 0.5 < MNSQ < 1.5 in question no. 10 it was not included because the value had not entered 0.5 which if it had a value of less than 0.5 then it would be a less productive question to use but did not reduce the quality of the question(Maulana et al., 2023; Sumintono & Widhiarso, 2014). while for question no. 9 it passes 1.5 which if it has a value of more than 1.5 then it will reduce the quality of the measurement system it self (Sumintono & Widhiarso, 2015). on the ZSTD outfit criteria, it is found that all questions meet the criteria, namely -2.0 <ZSTD< + 2.0 where in the data obtained the lowest value of ZSTD is 0.0 and the highest is 1.2 so the question meets the criteria. In the Pt mean Corr criterion, it was found that none of the questions met the criteria, namely 0.4 < Pt Measure Corr< 0.85, where in the data none of them met this value.

An instrument can be said to be valid if one of the following three criteria Outfit means square (MNSQ) 0.5<MNSQ<1.5, Outfit Z-standard (ZSTD) -2.0<ZSTD<+2.0, Point Measure Correlation (Pt mean Corr) 0.4< Pt Measure Corr<0.85. (Maulana et al., 2023; Sumintono & Widhiarso, 2014) the results of the analysis conducted by the validator show that all of the validation criteria above show that they meet all of the expected validation criteria so that it can be concluded that the test instrument has good content validity the results of the analysis conducted by the validator show that all of the validation criteria above show that they meet all of the expected validation criteria so that it can be concluded that the test instrument has good content validity and does not need any revision.

2) Empirical validity

Based on data analysis conducted at the first, third and fifth levels, it is found that all questions at the first, third and fifth levels are valid on all 15 questions that have been made. The following empirical validity data at the first level can be seen in table 5, the third level can be seen in table 6 and empirical validity data at the fifth level can be seen in table 7. Where each question still falls into one of the requirements of the following three things Outfit means square (MNSQ) 0.5 <MNSQ <1.5, Outfit Z-standard (ZSTD) -2.0 <ZSTD <+2.0, Point Measure Correlation (Pt mean Corr) 0.4 < Pt Measure Corr<0.85.

Table 5 Empirical validity at first level

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
8	3	32	2.03	.62	1.10	.36	2.61	1.74	A-.02	.22	90.6	90.6	S8
7	9	32	.57	.42	1.28	1.42	1.55	1.60	B .01	.34	65.6	74.0	S7
12	6	32	1.16	.48	1.14	.56	1.33	.81	C .12	.30	81.3	81.2	S12
9	16	32	-.51	.38	1.03	.24	1.25	1.33	D .31	.38	68.8	65.6	S9
11	8	32	.75	.43	1.09	.47	1.15	.53	E .22	.33	75.0	76.1	S11
15	11	32	.24	.40	1.08	.55	1.15	.65	F .25	.36	68.8	70.3	S15
13	12	32	.08	.39	1.09	.66	1.05	.28	G .28	.36	65.6	68.8	S13
14	8	32	.75	.43	1.00	.09	.97	.04	g .32	.33	81.3	76.1	S14
6	15	32	-.37	.38	.93	-.55	.87	-.66	f .46	.38	65.6	66.0	S6
1	24	32	-1.78	.44	.90	-.37	.82	-.48	e .46	.36	81.3	77.2	S1
3	20	32	-1.11	.39	.90	-.66	.81	-.85	d .50	.38	78.1	68.4	S3
4	12	32	.08	.39	.80	-1.36	.76	-1.08	c .56	.36	78.1	68.8	S4
2	25	32	-1.98	.46	.75	-.99	.61	-1.06	b .60	.35	84.4	79.8	S2
5	12	32	.08	.39	.75	-1.76	.69	-1.46	a .62	.36	78.1	68.8	S5
MEAN	12.1	32.0	.30	.52	.99	-.09	1.12	.10			75.9	73.7	
P.SD	6.9	.0	1.52	.35	.15	.85	.49	1.02			7.6	6.7	

In the MNSQ outfit criterion, it is found that there are several questions that do not enter the criteria, namely 0.5 < MNSQ < 1.5, namely in s8-1 and s12-1 so that these two questions have a value of more than 1.5 which if they have a value of more than 1.5 then it will reduce the quality of the measurement system itself (Istiyono, 2014; Sumintono & Widhiarso, 2015) Sumintono & Widhiarso, 2015). In the MNSQ outfit criteria, it was found that there were several questions that did not enter the criteria, namely 0.5 < MNSQ < 1.5, namely in s8 and s12 so that these two questions had a value of more than 1.5 which if they had a value of more than 1.5, it would reduce the quality of the measurement system itself (Sumintono & Widhiarso, 2015 & (Istiyono, 2014). In the ZSTD outfit criteriaIt is obtained that all questions meet the criteria, namely -2.0 < ZSTD < + 2.0 where in the data obtained the lowest value of ZSTD is -1.57 and the highest is 1.86 so all questions meet the criteria. In the Pt mean Corr criteria, it was found that there were several questions that did not meet the criteria, namely 0.4 < Pt Measure Corr < 0.85 so that in the data there were s8, s12, s1, s7, s13 and s11 which had a value of less than 0.4.

Table 6 Empirical validity at third level

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
12	6	32	.89	.49	1.04	.22	1.70	1.24	A .24	.35	87.1	82.7	S12
5	3	32	1.82	.64	1.29	.73	1.32	.62	B .08	.28	90.3	90.2	S5
15	12	32	-.26	.41	1.24	1.56	1.32	1.12	C .24	.42	61.3	68.6	S15
8	6	32	.89	.49	1.30	1.05	1.29	.67	D .16	.35	74.2	82.7	S8
1	16	32	-.90	.40	1.09	.64	1.27	1.12	E .37	.45	64.5	67.2	S1
10	5	32	1.15	.53	1.05	.25	1.14	.42	F .27	.33	83.9	85.1	S10
6	15	32	-.74	.40	1.09	.71	1.05	.27	G .39	.44	54.8	67.2	S6
7	4	32	1.45	.57	1.08	.33	.86	.05	H .27	.30	83.9	87.6	S7
9	17	32	-1.05	.40	.94	-.40	.84	-.64	g .52	.46	67.7	68.5	S9
13	8	32	.45	.45	.93	-.24	.78	-.43	f .44	.38	80.6	77.4	S13
2	25	32	-2.57	.50	.92	-.17	.78	-.27	e .55	.49	80.6	82.8	S2
11	10	32	.08	.42	.85	-.87	.79	-.56	d .50	.40	74.2	72.2	S11
4	15	32	-.74	.40	.78	-1.69	.69	-1.38	c .60	.44	74.2	67.2	S4
14	4	32	1.45	.57	.73	-.63	.60	-.38	b .47	.30	90.3	87.6	S14
3	22	32	-1.92	.44	.65	-1.77	.52	-1.50	a .72	.48	83.9	76.3	S3
MEAN	11.2	32.0	.00	.47	1.00	-.02	1.00	.02			76.8	77.6	
P.SD	6.7	.0	1.26	.08	.19	.92	.32	.83			10.4	8.2	

In the MNSQ outfit, it is found that all questions meet the criteria, namely 0.5 < MNSQ < 1.5. In the ZSTD outfit criteria, it is found that all questions meet the criteria, namely -2.0 <ZSTD <+2.0. On the Pt mean Corr criterion, it is found that there are several questions that do not meet the criteria, namely 0.4 < Pt Measure Corr < 0.85 in the data contained in s12, s5, s15, s8, s1, s10, s6 and s7 which have numbers less than 0.4.

Table 7 Empirical validity at fifth level

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
1	108	32	-.50	.18	2.20	3.55	2.19	3.40	A .44	.31	34.4	54.6	S1
4	102	32	-.29	.19	1.66	1.98	1.60	1.79	B .30	.30	53.1	63.0	S4
15	84	32	.34	.18	1.51	1.80	1.57	1.89	C .39	.30	43.8	54.4	S15
5	94	32	-.01	.19	1.40	1.28	1.43	1.32	D .35	.29	56.3	67.8	S5
3	100	32	-.22	.19	.91	-.21	.90	-.22	E .46	.30	71.9	67.6	S3
9	90	32	.13	.19	.84	-.50	.85	-.45	F .35	.30	62.5	62.8	S9
12	96	32	-.08	.19	.85	-.42	.82	-.50	G .20	.29	71.9	70.0	S12
13	90	32	.13	.19	.83	-.51	.83	-.50	H .35	.30	65.6	62.8	S13
2	98	32	-.15	.19	.80	-.62	.81	-.56	g .01	.30	68.8	63.4	S2
11	88	32	.20	.19	.78	-.77	.80	-.65	f .16	.30	65.6	62.3	S11
10	90	32	.13	.19	.72	-.99	.70	-1.02	e .09	.30	71.9	62.8	S10
6	92	32	.06	.19	.65	-1.23	.65	-1.23	d .49	.30	78.1	69.7	S6
8	96	32	-.08	.19	.60	-1.45	.61	-1.37	c .16	.29	78.1	70.0	S8
7	98	32	-.15	.19	.55	-1.69	.56	-1.60	b-.07	.30	71.9	63.4	S7
14	80	32	.47	.18	.55	-2.22	.52	-2.28	a .57	.31	43.8	35.7	S14
MEAN	93.7	32.0	.00	.19	.99	-.13	.99	-.13			62.5	62.0	
P.SD	6.9	.0	.24	.00	.46	1.53	.47	1.49			13.0	8.4	

Empirical validity data at the fifth level can be seen in table 11 In the MNSQ outfit criteria, it was found that there were several questions that did not enter the criteria, namely 0.5 < MNSQ < 1.5 so that the data for questions s1 and s4 had a value of more than 1.5 from 1.5, if it has a value of more than 1.5 then it will reduce the quality of the measurement system itself (Sumintono & Widhiarso, 2015) In the outfit ZSTD criterion, it is found that all questions meet the criteria, namely -2.0 <ZSTD< +2.0, where the data obtained is less than -2.0, namely in s14, which if the value is less than -2.0 then the data is too easy to predict and which passes +2.0, namely s1, if the value passes +2.0 it will produce data that is not expected according to its needs (Sumintono & Widhiarso. 2015). In the Pt mean Corr criteria, it was found that there were several questions that did not meet the criteria, namely 0.4 < Pt Measure Corr < 0.85 in the data contained in s4, s15, s5, s9, s12, s2, s13, s10, s8 and s7 which had a value of less than 0.4.

In each of the questions at this fifth level, the question still falls into one of the following three conditions Outfit means square (MNSQ) 0.5 < MNSQ < 1.5, Outfit Z-standard (ZSTD) -2.0

< ZSTD < + 2.0, Point Measure Correlation (Pt mean Corr) 0.4 < Pt Measure Corr < 0.85. The results of the Rasch model analysis are declared valid if they meet at least one of the three conditions (Maulana et al., 2023; Sumintono & Widhiarso, 2014).

It can be concluded that all questions at the first, third and fifth levels are valid because they have met at least one of the three conditions (Maulana et al., 2023; Sumintono & Widhiarso, 2014).

b. Reliability

The reliability of the instrument at the first level can be seen in table 8 with a value of 0.82, and the reliability at the third level can be seen in table 9 with a value of 0.84 with the same category, namely good, which means that the test instrument developed is classified as reliable so that the test instrument can be trusted and provides results that are not different if retested (Febriano et al., 2021).

Table 8 Reliability of items at first level

SUMMARY OF 15 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	12.9	32.0	.00	.43	.99	-.09	1.12	.10
SEM	1.7	.0	.29	.02	.04	.24	.14	.28
P.SD	6.3	.0	1.06	.06	.15	.85	.49	1.02
S.SD	6.5	.0	1.10	.06	.16	.88	.51	1.06
MAX.	25.0	32.0	2.03	.62	1.28	1.42	2.61	1.74
MIN.	3.0	32.0	-1.98	.38	.75	-1.76	.61	-1.46

REAL RMSE	.45	TRUE SD	.96	SEPARATION	2.15	Item	RELIABILITY	.82
MODEL RMSE	.43	TRUE SD	.97	SEPARATION	2.23	Item	RELIABILITY	.83
S.E. OF Item MEAN = .29								

Table 9 Item reliability at third level

SUMMARY OF 15 MEASURED (NON-EXTREME) Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	11.2	32.0	.00	.47	1.00	-.02	1.00	.02
SEM	1.8	.0	.34	.02	.05	.24	.09	.22
P.SD	6.7	.0	1.26	.08	.19	.92	.32	.83
S.SD	6.9	.0	1.31	.08	.20	.95	.33	.86
MAX.	25.0	32.0	1.82	.64	1.30	1.56	1.70	1.24
MIN.	3.0	32.0	-2.57	.40	.65	-1.77	.52	-1.50

REAL RMSE	.50	TRUE SD	1.16	SEPARATION	2.32	Item	RELIABILITY	.84
MODEL RMSE	.48	TRUE SD	1.17	SEPARATION	2.44	Item	RELIABILITY	.86
S.E. OF Item MEAN = .34								

Reliability at the fifth level which can be seen in table 10 with a value of 0.70 with a sufficient category so that the test instrument is also classified as reliable and trustworthy (Febriano et al., 2021).

Table 10 Item reliability at fifth level

SUMMARY OF 15 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	93.7	32.0	.00	.19	.99	-.13	.99	-.13
SEM	1.9	.0	.06	.00	.12	.41	.12	.40
P.SD	6.9	.0	.24	.00	.46	1.53	.47	1.49
S.SD	7.2	.0	.25	.00	.48	1.59	.48	1.55
MAX.	108.0	32.0	.47	.19	2.20	3.55	2.19	3.40
MIN.	80.0	32.0	-.50	.18	.55	-2.22	.52	-2.28

REAL RMSE	.20	TRUE SD	.13	SEPARATION	1,80 Item	RELIABILITY .70
MODEL RMSE	.19	TRUE SD	.15	SEPARATION	1.92 Item	RELIABILITY .74
S.E. OF Item MEAN = .06						

Item RAW SCORE-TO-MEASURE CORRELATION = -1.00
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000

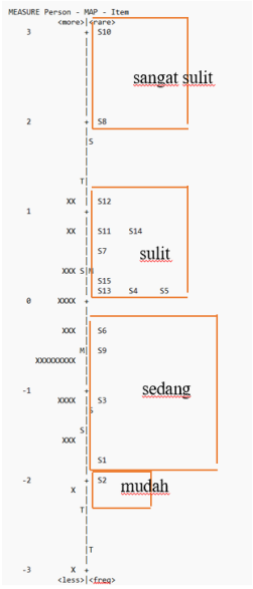
c. Difficulty index

Based on the results of the data analysis conducted, the first and third levels are in table 9. The fifth level can be seen in table 11.

Table 11 Question difficulty index at first level

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
10	0	32	4.48	1.82	MAXIMUM MEASURE				.00	.00	100.0	100.0	S10
8	3	32	2.03	.62	1.10	.36	2.61	1.74	-.02	.22	90.6	90.6	S8
12	6	32	1.16	.48	1.14	.56	1.33	.81	.12	.30	81.3	81.2	S12
11	8	32	.75	.43	1.09	.47	1.15	.53	.22	.33	75.0	76.1	S11
14	8	32	.75	.43	1.00	.09	.97	.04	.32	.33	81.3	76.1	S14
7	9	32	.57	.42	1.28	1.42	1.55	1.60	.01	.34	65.6	74.0	S7
15	11	32	.24	.40	1.08	.55	1.15	.65	.25	.36	68.8	70.3	S15
4	12	32	.08	.39	.80	-1.36	.76	-1.08	.56	.36	78.1	68.8	S4
5	12	32	.08	.39	.75	-1.76	.69	-1.46	.62	.36	78.1	68.8	S5
13	12	32	.08	.39	1.09	.66	1.05	.28	.28	.36	65.6	68.8	S13
6	15	32	-.37	.38	.93	-.55	.87	-.66	.46	.38	65.6	66.0	S6
9	16	32	-.51	.38	1.03	.24	1.25	1.33	.31	.38	68.8	65.6	S9
3	20	32	-1.11	.39	.90	-.66	.81	-.85	.50	.38	78.1	68.4	S3
1	24	32	-1.78	.44	.90	-.37	.82	-.48	.46	.36	81.3	77.2	S1
2	25	32	-1.98	.46	.75	-.99	.61	-1.06	.60	.35	84.4	79.8	S2
MEAN	12.1	32.0	.30	.52	.99	-.09	1.12	.10			75.9	73.7	
P.SD	6.9	.0	1.52	.35	.15	.85	.49	1.02			7.6	6.7	

Table 12 wright map of question difficulty index at first level

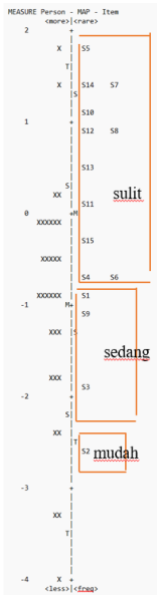


Based on what can be seen in table 11, the index of difficulty at the first level, the index of difficulty is sorted from highest to lowest. The highest difficulty index is found in question no. 10 at the first level where the difficulty value reaches a maximum which means that at s10-1, students who have tried to work on the question cannot answer correctly, so that at s10 it is a very difficult question level and cannot be reached by students with high ability students though (Safihin et al., 2019) the lowest difficulty index is in question no. 2 with a value of -1.98 with a low difficulty level. It can be seen in table 12 of the wright map of the problem difficulty index that the highest logit value is 3 and the lowest is -2, where the group of items that can be reached by the ability of students lies at a logit value of -2 to 2 (Safihin et al., 2019). The difficulty index for a good test is a test that has a varying difficulty index with 4 categories of question difficulty index, namely from easy, medium, difficult and very difficult.

Table 13 question difficulty index at third level

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
5	3	32	1.82	.64	1.29	.73	1.32	.62	.08	.28	90.3	90.2	S5
7	4	32	1.45	.57	1.08	.33	.86	.05	.27	.30	83.9	87.6	S7
14	4	32	1.45	.57	.73	-.63	.60	-.38	.47	.30	90.3	87.6	S14
10	5	32	1.15	.53	1.05	.25	1.14	.42	.27	.33	83.9	85.1	S10
8	6	32	.89	.49	1.30	1.05	1.29	.67	.16	.35	74.2	82.7	S8
12	6	32	.89	.49	1.04	.22	1.70	1.24	.24	.35	87.1	82.7	S12
13	8	32	.45	.45	.93	-.24	.78	-.43	.44	.38	80.6	77.4	S13
11	10	32	.08	.42	.85	-.87	.79	-.56	.50	.40	74.2	72.2	S11
15	12	32	-.26	.41	1.24	1.56	1.32	1.12	.24	.42	61.3	68.6	S15
4	15	32	-.74	.40	.78	-1.69	.69	-1.38	.60	.44	74.2	67.2	S4
6	15	32	-.74	.40	1.09	.71	1.05	.27	.39	.44	54.8	67.2	S6
1	16	32	-.90	.40	1.09	.64	1.27	1.12	.37	.45	64.5	67.2	S1
9	17	32	-1.05	.40	.94	-.40	.84	-.64	.52	.46	67.7	68.5	S9
3	22	32	-1.92	.44	.65	-1.77	.52	-1.50	.72	.48	83.9	76.3	S3
2	25	32	-2.57	.50	.92	-.17	.78	-.27	.55	.49	80.6	82.8	S2
MEAN	11.2	32.0	.00	.47	1.00	-.02	1.00	.02			76.8	77.6	
P.SD	6.7	.0	1.26	.08	.19	.92	.32	.83			10.4	8.2	

Table 14 wright map of question difficulty index at third level

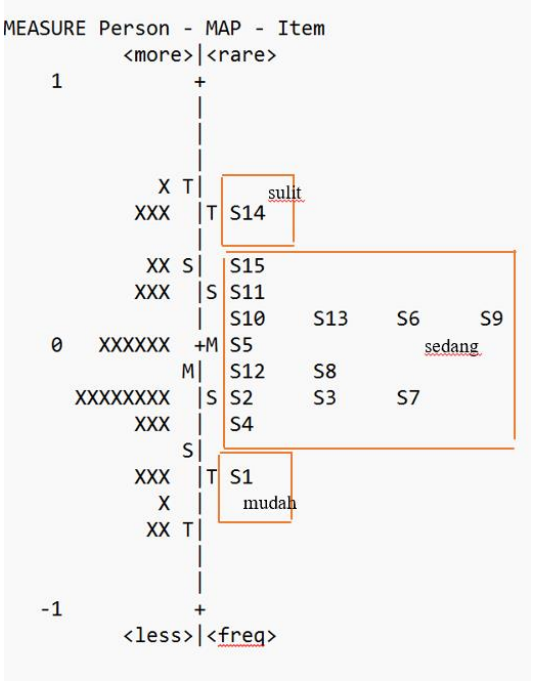


The results of the question difficulty index at the third level can be seen in table 10, there is the highest difficulty index, namely at s5 with a value of 1.82 and the lowest difficulty index is at s2 with a value of -2.57 which is where this question is included in the not too difficult and not too easy so that the question can carry out its function properly (Sumintono & Widhiarso, 2015. It can be seen in table 11 that the question difficulty index is divided into 3, namely, easy, medium and difficult.

Table 15 question difficulty index at fifth level

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
14	80	32	.47	.18	.55	-2.22	.52	-2.28	.57	.31	43.8	35.7	S14
15	84	32	.34	.18	1.51	1.80	1.57	1.89	.39	.30	43.8	54.4	S15
11	88	32	.20	.19	.78	-.77	.80	-.65	.16	.30	65.6	62.3	S11
9	90	32	.13	.19	.84	-.50	.85	-.45	.35	.30	62.5	62.8	S9
10	90	32	.13	.19	.72	-.99	.70	-1.02	.09	.30	71.9	62.8	S10
13	90	32	.13	.19	.83	-.51	.83	-.50	.35	.30	65.6	62.8	S13
6	92	32	.06	.19	.65	-1.23	.65	-1.23	.49	.30	78.1	69.7	S6
5	94	32	-.01	.19	1.40	1.28	1.43	1.32	.35	.29	56.3	67.8	S5
8	96	32	-.08	.19	.60	-1.45	.61	-1.37	.16	.29	78.1	70.0	S8
12	96	32	-.08	.19	.85	-.42	.82	-.50	.20	.29	71.9	70.0	S12
2	98	32	-.15	.19	.80	-.62	.81	-.56	.01	.30	68.8	63.4	S2
7	98	32	-.15	.19	.55	-1.69	.56	-1.60	-.07	.30	71.9	63.4	S7
3	100	32	-.22	.19	.91	-.21	.90	-.22	.46	.30	71.9	67.6	S3
4	102	32	-.29	.19	1.66	1.98	1.60	1.79	.30	.30	53.1	63.0	S4
1	108	32	-.50	.18	2.20	3.55	2.19	3.40	.44	.31	34.4	54.6	S1
MEAN	93.7	32.0	.00	.19	.99	-.13	.99	-.13			62.5	62.0	
P.SD	6.9	.0	.24	.00	.46	1.53	.47	1.49			13.0	8.4	

Table 16 wright map of fifth level question difficulty index



Based on what can be seen in table 11, the difficulty index at the fifth level, the difficulty index is sorted based on the highest to the lowest, so that the highest difficulty index is in question no. 14 with a value of 0.47 and the lowest difficulty index is in s1 with a value of -0.5 which is a question that is not too difficult and not too easy so that the question can carry out its function properly (Sumintono & Widhiarso2015. It can be seen in table 17 that the difficulty index of the question is divided into 3, namely, easy, medium and difficult.

d. Differentiating power

The differentiation of questions on each item can be seen from the separation value and entered into the formula, namely,

$$H = \frac{[(4\ x\ \text{SEPARATION}\ ) + 1]}{3}$$

Where the H value is obtained at the first, third and fifth levels, the H value obtained is 3, by getting a value of 3, the differentiating power of a question can be categorized as good, and

has 3 distribution categories, namely difficult, medium and easy (Sumintono & Widhiarso, Aplikasi PemodelanRasch pada Assessment Pendidikan, 2015).

CONCLUSION

Based on the research conducted, this study has successfully developed a five-tier diagnostic test instrument that shows good validity. The reliability of the instrument at the first and third levels reached 0.86 (very good), while at the fifth level it was 0.70 (quite good). The index of difficulty and differentiation of questions on this instrument also varied, indicating the instrument's ability to distinguish students with different abilities.

REFERENCES

Agustin, U., Susilaningsih, E., & Nurhayati, S. (2022). Chemistry In Education. *Chemistry In Education*, *1*(11), 133–139. Https://Doi.Org/10.1111/J.1949-8594.1902.Tb00418.X

Fajriyyah, N. S., & Ermawati, F. U. (2020). The Validity And Reliability Of Five-Tier Conception Diagnostic Test For Kinetic Theory Of Gases. *IPF : Inovasi Pendidikan Fisika*, *09*(02), 126–132.

Febriano, R., Tandililing, E., & Enawaty, E. (2021). Pengembangan Instrumen Tes Kemampuan Berpikir Kritis Matematis Dengan Menggunakan Analisis Model Rasch Pada Siswa SMP. *Jurnal Pendidikan Dan Pembelajaran Khatulistiwa*, *10*(9), 1–12.

Istiyono, E. (2014). DEVELOPING HIGHER ORDER THINKING SKILL TEST OF PHYSICS (Physthots) FOR SENIOR HIGH SCHOOL STUDENTS. *Penelitian Dan Evaluasi Pendidikan*, *18*(1), 1–12. Http://Journal.Uny.Ac.Id/Index.Php/Jpep/Article/View/2120

Lailiyah, S., & Ermawati, F. U. (2020). Materi Gelombang Bunyi: Pengembangan Tes Diagnostik Konsepsi Berformat Five-Tier, Uji Validitas Dan Reliabilitas Serta Uji Terbatas. *Jurnal Pendidikan Fisikatadulakoonline (JPFT)*, *8*(3), 104–119.

Liu, X. (2012). Using And Developing Measurement Instruments In Science Education: A Rasch Modeling Approach. In *Science Education* (Vol. 96, Issue 1). Https://Doi.Org/10.1002/Sce.20477

Maulana, S., Rusilowati, A., Nugroho, S. E., & Susilaningsih, E. (2023). Implementasi Rasch Model Dalam Pengembangan Instrumen Tes Diagnostik. *Prosiding Seminar Nasional Pascasarjana (PROSNAMPAS)*, *6*(1), 748–757. Https://Proceeding.Unnes.Ac.Id/Index.Php/Snpasca/Article/View/2214

Monita, Ade, F., & Bambang, S. (2016). *IDENTIFIKASI DAN ANALISIS MISKONSEPSI SISWA MENGGUNAKAN THREE-TIER MULTIPLE CHOICE DIAGNOSTIC INSTRUMENT PADA KONSEP KESETIMBANGAN KIMIA*. *7*(1), 1–16.

Nisa, K., & Sudrajat, A. (2023). Pengembangan Instrumen Tes Diagnostik Five-Tier Untuk Mengidentifikasi Miskonsepsi Siswa Kelas XI Pada Materi Laju Reaksi. *PENDIPA Journal Of Science Education*, *7*(2), 127–136. Https://Doi.Org/10.33369/Pendipa.7.2.127-136

Permatasari, M. B., Muchson, M., Hakimah, N., Rokhim, D. A., Herunata, H., & Yahmin, M. (2022). Identifikasi Miskonsepsi Materi Kesetimbangan Kimia Pada Siswa SMA Menggunakan Tes Three Tier Berbasis Web. *Jurnal Inovasi Pendidikan Kimia*, *16*(1), 1–7. Https://Doi.Org/10.15294/Jipk.V16i1.29407

Putri, W. K., & Ermawati, F. U. (2021). Pengembangan, Uji Validitas Dan Reliabilitas Tes Diagnostik Five-Tier Untuk Materi Getaran Harmonis Sederhana Beserta Hasil Uji Coba. *PENDIPA Journal Of Science Education*, *5*(1), 92–101. Https://Doi.Org/10.33369/Pendipa.5.1.92-101

Safihin, M., S, H. T. M., & Hamdani. (2019). Pengembangan Tes Menggunakan Model Rasch. *Jurnal Pendidikan Dan Pembelajaran Khatulistiwa*, *8*(6).

Setiawan, D., & Faoziyah, N. (2020). Development Of A Five-Tier Diagnostic Test To Reveal The Student Concept In Fluids. *Physics Communication*, *4*(1), 6–13. Http://Journal.Unnes.Ac.Id/Nju/Index.Php/Pc

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

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelanrasch Pada Assessment Pendidikan.*

Zulfadli, & Munawwarah, I. (2016). Identifikasi Pemahaman Siswa Terhadap Konsep Kelarutan Dan Hasil Kali Kelarutan Dengan Menggunakan Tes Diagnostik Three-Tier Multiple Choice. *Edukasi Kimia*, *1*(1), 32–40.