J. Pijar MIPA, Vol. 18 No. 2, March 2023: 183-193 ISSN 1907-1744 (Cetak)
DOI: 10.29303/jpm.v18i2.4721 ISSN 2460-1500 (Online)

DEVELOPMENT OF A TEST INSTRUMENT USING MULTIPLE REPRESENTATIONS TO ASSESS STUDENTS UNDERSTANDING: A RASCH MODEL ANALYSIS

Yayang Andita Putri and Faizah Qurrata Aini*
Chemistry Education Study Program, Faculty of Mathematics and Natural Sciences, Universitas Negeri Padang, West Sumatera, Indonesia
*Email: faizah_qurrata@fmipa.unp.ac.id

Received: January 28, 2023. Accepted: March 6, 2023. Published: March 30, 2023

Abstract: Today's test instruments used in schools are still ineffective in assessing students' multi-representational understanding of the "rate of reaction" topic. Hence, teachers need to accurately and comprehensively know the depth of students' understanding. In this study, a test instrument was produced to assess the level of macroscopic, submicroscopic, and symbolic understanding of high school students in the rates of reaction topic. The method used is Research and Development (R&D) based on the Rasch modeling approach, which is modified in ten stages, i.e., (1) defining construct, (2) identifying question indicators, (3) compiling items, scoring rubrics, scoring guidelines, and the guidelines to analyze student's level of understanding, (4) conducting the pilot test, (5) analyzing data with the Rasch model, (6) reviewing the suitability of items, (7) reviewing the Wright map, (8) repeating steps 4-7 until all items fit, (9) claiming the instrument's quality, and (10) documenting the instrument. All items proved valid because they met the fit criteria of MNSQ, ZSTD, and PT Measure scores. Instrument reliability reaches the score of 0.89 (good). The instrument has a four-difficulty index, dominated by moderate difficulty (70.6%). This instrument is relatively easy. The instrument also has four discrimination indexes: very difficult, difficult, medium, and easy. It proves that this assessment instrument is of good quality and suitable for school use.

Keywords: *Instrument Test, Multiple Representation, Rates of Reaction, Rasch Model.*

INTRODUCTION

As a science with the majority of abstract matters that are difficult to observe directly, the process of conveying chemistry concepts to students has been a challenge teachers face for years. Three levels of multiple representations are being used (macroscopic, sub-microscopic, and symbolic) in the learning process in schools. The multiple representations could help students to overcome the abstractness of chemistry concepts in different depths of understanding level, so the student can learn easier and appreciate chemistry better [1]. These levels cannot stand alone as they complement and are interconnected. The macroscopic level is a natural phenomenon that can be observed directly. The sub-microscopic level describes said phenomenon on the molecular level, and the symbolic level represents it quantitatively using formulas, equations, mathematical operations, and graphs [2]. The application of multiple representations is needed to form a complete mental model for understanding chemistry and avoiding the formation of misconceptions [3,4].

The rate of reaction is a branch of chemistry that studies the reaction speed. To explain the reasoning behind the rate differences of various reactions, students must understand the process behind altering reactants into products at the molecular level and the laws and calculations behind those reactions [5]. Even though teachers from two schools in Padang City and Padang Pariaman Regency have already implemented the three levels of representation (macroscopic, sub-microscopic, and symbolic) while teaching the rate of reaction, the test instrument used is still not effective in assessing students' understanding of multi-representation in the rate of reaction topic. From observations done on 54 students in said schools, 96% of students stated that they feel capable of answering questions that required strong understanding at the macroscopic level of the topic, 74% at the sub-microscopic level, and 85% at the symbolic level. However, school test instruments focus only on the symbolic and/or macroscopic levels. Meanwhile, the test instrument to assess students' understanding of the sub-microscopic level still needs to be created, when in fact, interconnection in the test instruments is also very necessary so that teachers can find out students' understanding and mental models accurately and comprehensively.

The test instruments developed in this study are arranged in essay form. The form was chosen because it can require students to express their understanding in their own words, reduce the opportunity for lucky guess answers, and can show the level of students' understanding of the problems asked more accurately [6]. Each item is designed with several sub-items representing each macroscopic, sub-microscopic, and symbolic level that is interconnected by one another. Thus, it can assess students' level of understanding in the rate of reaction topic.

This study aimed to produce a test instrument to assess students' understanding of the macroscopic, sub-microscopic, and symbolic

levels in the rates of reaction topic. The instrument's feasibility needs to be proven for its quality through the following components: validity, reliability, discriminatory index, and item difficulty index [7]. Rasch modeling is used as the reference to measure the instrument quality. The model was chosen because of its advantages in providing complete and accurate information about the items and the abilities of the subjects involved in this test. Rasch's model computes the latent construct of the issues and their relation to the item difficulty. It will produce an interval level scale known as logit [8,9]. The logit (logarithm odds unit) produces a measurement scale with equal intervals. People and items are placed equally in a continuous line according to their respective skill/difficulty levels [10,11]. This logit transformation follows a normal distribution curve, which means it has test-free and person-free characteristics [12]. Therefore, the measurement of the Rasch model is more valid because it calibrates three aspects at once, i.e., the measurement scale, students' abilities, and the test items. Rasch modeling also allows researchers to predict the best score for missing data, identify students' abilities and error responses, find out if there are guessed answers, detect subjects that are inappropriate and need to be removed from the sample, and detect items that need to be revised or deleted [10].

RESEARCH METHOD

This research development (R&D) refers to the Rasch Modeling Approach, which was adapted and modified from research by Wei et al. [13]. The resulting product is six essay test instruments to test understanding of the macroscopic, sub-microscopic, and symbolic levels of high school students on the material of reaction rate. Five chemistry lecturers carried out proof of the validity of the contents of the instrument at FMIPA UNP. The instrument used was a questionnaire to prove validity using four Likert scales, namely strongly disagree (STS), disagree (TS), agree (S), and strongly agree (SS). The data obtained were then analyzed using Minifac (Facets) 3.84.1 software. After its validity was proven, the instrument was tested on subjects, namely 30 students in class XI MIPA 2 SMA Negeri 1 Lubuk Alung in the 2022/2023 Academic Year. The selection of the number of issues is based on the recommendation of the Rasch model with a confidence level of 95% [10]. The object of this study is the quality of the instrument in terms of validity, reliability, difficulty index, and item discriminating power. Data from the trial results were analyzed with the Rasch model using Ministep 5.2.4.0 software.

The stages of developing test instruments based on the Rasch Modeling Approach are divided into ten procedures, namely: (1) defining the constructs of the questions in the learning progression, (2) identifying the question indicators, (3) compiling item items, scoring rubrics, scoring guidelines, and understanding level analysis guidelines multi representation of students, (4) conducting instrument trials on subjects, (5) applying the Rasch model to raw data using Ministep software, (6) reviewing the suitability of items based on the Rasch model and revising items if necessary, (7) reviewing Wright maps and adding or reduce the items if needed, (8) repeating steps 4-7 until all items fit, (9) determine validity, reliability, difficulty index, and differentiability claims for each item, and (10) document the instrument [13]. The valid instrument has items with outfit scores of $0.5 < MNSQ < 1.5$, $-2.0 < ZSTD < +2.0$, and $0.4 < Pt$ measure < 0.85. The reliable instrument has a Cronbach Alpha score > 0.7 and item reliability > 0.80. A good difficulty index is shown by an even distribution of items on the Wright Map, while instruments with a good discrimination index have a separation score of 2.0 to 3.0 [10,14].

RESULTS AND DISCUSSION

In this study, a test instrument that consists of 6 questions was produced to assess students' understanding of the macroscopic, sub-microscopic, and symbolic levels in the rates of reaction topic. Each item contains 2 or 3 sub-items interconnected at every representation level regarding the constructs needed for each material. It makes the total of sub-items in this instrument 17 items. The development of this instrument was carried out in 10 stages [13], which get the following results.

Defining construct

At this stage, the construct of the rates of reaction topic is identified into a learning progression based on Basic Competencies (KD) 3.6 and 3.7. The two basic competencies (KD) were generated as indicators of Competence Achievement (IPK), which contain the cognitive and representational levels, as seen in Table 1.

Identifying question indicators

One indicator question per Key IPK is generated based on the construct's definition above. Only the Key IPK is developed into the question indicators, containing the equivalent operational verbs to the KD. Meanwhile, the KKO on Supporting IPK has a level below it [15]. Hence, if students achieve the Key IPK, all Supporting IPKs are automatically achieved. The main topics tested for each question indicator are as follows: (1) the effect of concentration on the rate of reaction, (2) the effect of surface area on the rate of reaction, (3) the effect of temperature on the rate of reaction, (4) the effect of catalyst on the rate of reaction, (5) the order of the reaction, and (6) the rate constant for the reaction.

Table 1. Learning Progression

Basic Competencies (KD)	Key IPK	Cognitive Level	Representation
3.6 Explaining the factors that affect the rate of reaction using collision theory.	3.6.1 Explaining the effect of concentration on reaction rates using collision theory.	C2	*Macroscopic*, the observable form of reactants and products, concentration, reaction time. *Sub-microscopic is* the illustration of collisions between particles during the reaction. *Symbolic*, the rates of reaction curve.
	3.6.2 Explaining surface area's effect on reaction rates using collision theory.	C2	*Macroscopic is* the observable surface area of the reactants and reaction time. *Sub-microscopic is* the illustration of collisions between particles during the reaction. *Symbolic*, the rates of reaction curve.
	3.6.3 Explaining the effect of temperature on reaction rates using collision theory.	C2	*Macroscopic*, the observable form of reactants and products, temperature, and reaction time. *Sub-microscopic is* the illustration of collisions between particles during the reaction. *Symbolic*, the rates of reaction curve.
	3.6.4 Explaining the catalyst's effect on the reaction rate using the collision theory.	C2	*Macroscopic is* the observable form of reactants, products, and catalysts; reaction time. *Sub-microscopic* is the molecular illustration of the effect of the catalyst on the reaction. *Symbolic is an* activation energy curve.
3.7 Determining the reaction order and the rate constant based on the experimental data.	3.7.2 Determining the order of the reaction based on experimental data.	C3	*Macroscopic* concentration is the time needed to change the reactants into products (in some reactions, it can be observed through color changes). *Symbolic*, *the* reaction order formula, and the reaction order curve.
	3.7.3 Determining the rate constant of a reaction based on experimental data.	C3	*Macroscopic* concentration is the time needed to change the reactants into products. *Symbolic is* the reaction rate constant formula.

Developing questions and assessment rubrics

From the previous indicators, six essay items (with a total of 17 sub-items) were produced. Each of which had several sub-items (a, b, and c) that represented a level of representation (macroscopic, sub-microscopic, and symbolic) according to the needs of the IPK. Each representation level is interconnected to comprehensively test students' understanding and mental models [16,17]. The essay form was chosen because it requires students to think orderly and systematically, then compose ideas and express them in their language so that their understanding of the topic can be reflected better [18,19]. The following question is an example to give a better context of the interconnection on each representation level.

Question 1:
A researcher wants to observe the reaction rate between aluminum sheets (Al) and two sodium hydroxide solutions (NaOH) variations. The first experiment used 250 ml of 1 M NaOH solution. The second experiment used 250 ml of 3 M NaOH solution. The mass of aluminum added to each solution was 5 grams. Observe what happens 5 minutes later.

ISSN 1907-1744 (Cetak)
ISSN 2460-1500 (Online)

Figure. 1
Aluminum + 1 M NaOH

Figure. 2
Aluminum + 3 M NaOH

The reaction takes place in the following equation:

$$4Al(s) + 2NaOH(aq) \rightarrow 2Na^{+}(aq) + 4AlO^{2-}(aq) + 3H_2(g)$$

Based on the discourse above, answer the questions below!

a. Which factors affect the rate of the reaction above?
b. Choose the correct reaction rate curve for each of the reactions above! Explain your reasonings!

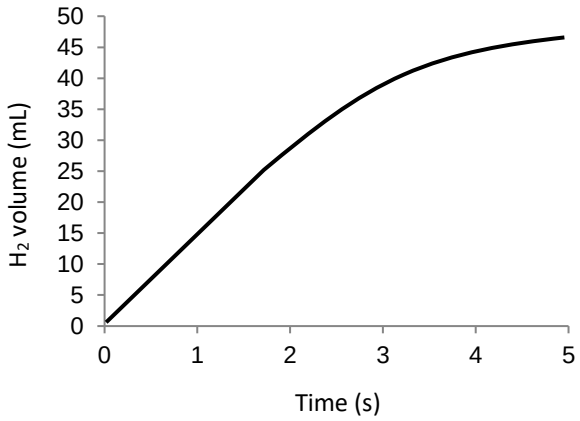


Figure 3. Curve 1

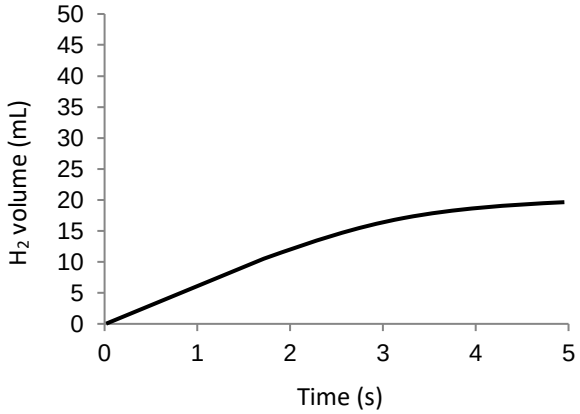


Figure 4. Curve 2

c. Figure 5 and Figure 6 illustrate the condition of the reactant particles in each experiment before the reaction occurs.

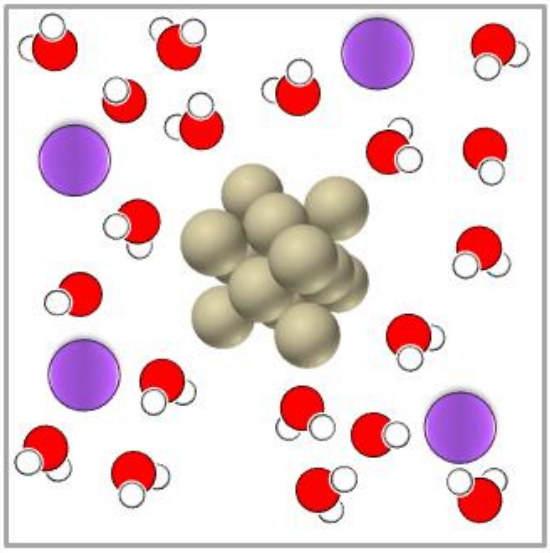


Figure. 5 Illustration 1

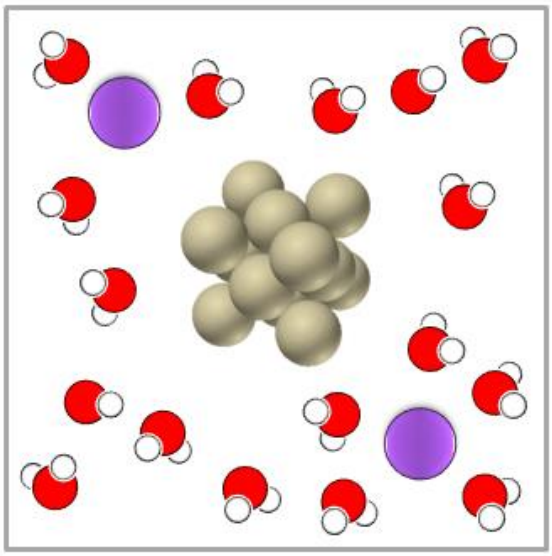
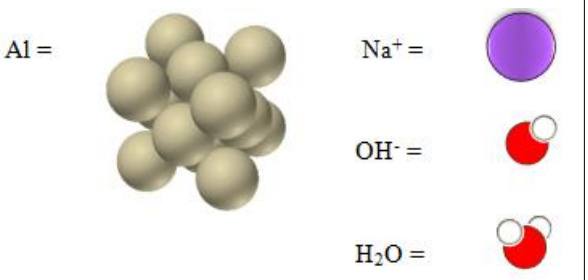


Figure 6. Illustration 2

Notes:
In solid, Al element forms a metallic crystal.



From the illustration above, choose which state of the reactant particles is suitable to describe the addition of 1 M and 3 M NaOH. Explain the reasons for your answer using the collision theory.

To avoid subjectivity while reviewing students' answers, an assessment rubric was prepared with answer keys, scoring guidelines, and a guideline for analyzing students' understanding levels of multi-representation in the rates of reaction topic [20]. The classification levels can be seen in Table 2 [21].

Table 2. The levels of students' multi-representation understanding in the rates of reaction topic

Level 3	Students can answer all macroscopic-symbolic-submicroscopic questions correctly. or Students can answer macroscopic-submicroscopic questions correctly.
Level 2	Students can answer the symbolic-submicroscopic questions correctly.
Level 1	Students can answer macroscopic-symbolic questions correctly. Or Students can correctly answer one of the macroscopic, submicroscopic, *or* symbolic questions.

Before being tested empirically on students, the content validity of the instrument must be verified first. Proof of validity is done to ensure that the contents of the instrument are representative, appropriate to the purpose of measurement, and can measure the expected construct [22,23]. A total of five Chemistry lecturers at Universitas Negeri Padang were selected as the subject matter expert and the expert media validators. All validators have given their rating in the form of Likert scales and agree that this instrument meets the criteria needed, which: (1) the items match the question indicators, (2) the scope to be measured is clear, (3) it has interconnection of the three levels of representation, (4) is capable of measuring the students' understanding of the multi representation in this topic, (5) question or commands are already demand answers appropriately, (6) items are clear and unambiguous, (7), illustrations are easy to read, (8) written in good Indonesian language, (9) do not contain elements of SARAPPPK, and (10) the answers provided have been proven by the scientific literature. The Likert scale was chosen as the most suitable scale to be analyzed with Rasch's Partial Credit Model (PCM) [24]. Post-analyze data of this content validity verification can be seen in the following table:

Table 3. The analysis results of the content validity with Subject Matter Experts

Strata Value	Reliability	Exact Agreements	Expected Agreements
7.82	0.97	58.0%	59.0%

The strata value (7.82) and reliability (0.97) scores show that the content validity with SMEs is proven to have very good reliability. Meanwhile, the exact agreement and expected agreement have a very small difference in value (1.0%), indicating that the analysis between the model and the exact data is classified as fit [25,26]. It can be concluded that the items produced do not need further revision and can proceed to be tested on the subject.

Conducting trials (pilot tests)

The purpose of the pilot test is to obtain the raw data to determine the instrument's validity, reliability, difficulty, and discrimination index. The pilot test was carried out on the subject, which is 30 students of XI MIPA at SMA Negeri 1 Lubuk Alung. The selection of subject classes was based on teacher recommendations which stated that XI MIPA 2 is a class with high enthusiasm for learning, and the students each have high, medium, and low abilities. In addition, the class has studied the rates of reaction topics at the end of the odd semester of the 2022/2023 Academic Year. The pilot test was carried out at the beginning of the even semester of the 2022/2023 Academic Year. Thus, this class meets the required criteria to be the subject of this trial based on the Rasch model [10].

Analyzing data with the Rasch model

The analysis performed on the raw data was divided into four types, which are: (1) validity, (2) reliability, (3) difficulty index, and (4) discrimination index. All of those analyzes were performed using Ministep 5.2.4.0 software. The results obtained include the following:

1. Validity

Validity analysis was performed using the Output Table: Item Fit Order menu. Each item must meet at least two of the following three criteria to be said as valid, i.e., the MNSQ, ZSTD, and Pt Measure outfit score [10]. If only one of the three criteria is successfully met, then the item cannot be used because it does not fit the model [27]. If it does not fit with the model, the best thing to do is to reformulate the item and repeat the fit analysis until the criterion is met [28]. The misfit items probably come from lucky guessing, reckless answers, scoring errors, etc. [12]. The results of this analysis obtained can be seen in Figure 7.

ISSN 1907-1744 (Cetak)
ISSN 2460-1500 (Online)

Item STATISTICS: MISFIT ORDER

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	G
2	40	30	1.52	.38	1.53	2.05	1.48	1.80	A .40	.51	62.1	67.7	1B	B
3	54	30	-.86	.50	1.51	1.46	1.30	.67	B .30	.38	82.8	82.6	1C	B
16	53	28	-1.73	.67	1.51	1.08	.65	-.11	C .40	.34	88.9	89.4	6A	B
9	37	30	1.96	.38	1.31	1.17	1.34	1.25	D .48	.54	62.1	70.4	3C	B
17	54	28	-2.25	.79	.96	.13	1.21	.57	E .27	.29	92.6	92.6	6B	B
14	28	29	-1.09	1.06	1.14	.44	.66	.20	F .17	.21	96.4	96.4	5B	A
12	57	29	-3.12	1.06	1.13	.43	.66	.20	G .17	.21	96.4	96.4	4C	B
13	57	29	-3.12	1.06	1.13	.43	.66	.20	H .17	.21	96.4	96.4	5A	B
15	54	29	-1.40	.59	.79	-.41	1.08	.36	I .41	.35	92.9	87.4	5C	B
6	42	30	1.23	.38	.93	-.28	.99	.04	h .40	.50	69.0	65.4	2C	B
5	51	30	-.21	.44	.97	-.03	.90	-.14	g .41	.42	75.9	74.8	2B	B
10	29	29	3.03	.40	.86	-.42	.86	-.39	f .44	.58	75.0	72.2	4A	B
7	4	30	5.48	.65	.76	-.43	.40	-.50	e .64	.50	89.7	89.6	3A	A
11	39	29	1.48	.39	.67	-1.56	.65	-1.55	d .56	.52	82.1	67.4	4B	B
1	28	30	-.30	.78	.53	-.80	.16	-.73	c .54	.27	93.1	93.1	1A	A
4	28	30	-.30	.78	.53	-.80	.16	-.73	b .54	.27	93.1	93.1	2A	A
8	28	30	-.30	.78	.53	-.80	.16	-.73	a .54	.27	93.1	93.1	3B	A
MEAN	40.2	29.4	.00	.65	.99	.10	.78	.02			84.8	84.0		
P.SD	14.3	.7	2.18	.24	.33	.92	.41	.78			11.5	11.3		

Figure 7. Validity test results with Item Fit Order

Based on the table above, items 1A, 2A, 3A, and 3B tend to be unfit because they have an MNSQ score < 0.5 (underfit). It means the item is hard to predict in the measurement [29]. However, the four items are worth keeping because the ZSTD and Pt Measure scores meet the criteria. Likewise, items 1C, 4C, 5A, 5B, and 6B tend not to fit because they have a Pt Measure score < 0.4, which is also worth keeping as the other two criteria fit. These items also do not have to be revised because they have positive Pt Measure scores, indicating no misleading. If the Pt Measure score is negative, then the item must be revised, as students with low abilities can answer difficult questions correctly, while students with high abilities answer wrongly [30].

Meanwhile, all items achieved the z score in the range of -2.0 < ZSTD < +2.0. Because the ZSTD acts as a t-test for the fit data hypothesis, this score implies that all items have a logical approximation of the data [10]. From this analysis, it can be concluded that all items in this instrument are valid.

2. Reliability

Reliability analysis was performed using the Output Table: Summary Statistics menu. Two components must be considered in determining the reliability of the instrument: the Cronbach Alpha score and Item Reliability.

Person RAW SCORE-TO-MEASURE CORRELATION = .76
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .86 SEM = 1.40
STANDARDIZED (50 ITEM) RELIABILITY = .86

SUMMARY OF 17 MEASURED (NON-EXTREME) Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	40.2	29.4	.00	.65	.99	.10	.78	.02
SEM	3.6	.2	.55	.06	.08	.23	.10	.19
P.SD	14.3	.7	2.18	.24	.33	.92	.41	.78
S.SD	14.7	.7	2.25	.25	.34	.94	.42	.80
MAX.	57.0	30.0	5.48	1.06	1.53	2.05	1.48	1.80
MIN.	4.0	28.0	-3.12	.38	.53	-1.56	.16	-1.55

REAL RMSE	.74	TRUE SD	2.05	SEPARATION	2.79	Item RELIABILITY .89
MODEL RMSE	.70	TRUE SD	2.07	SEPARATION	2.97	Item RELIABILITY .90
S.E. OF Item MEAN = .55						

Figure 8. Reliability test results with Summary Statistics

The results of the reliability analysis can be observed in Figure 8. The Cronbach Alpha score for this instrument is 0.86, which indicates that the interaction between students and the items is overall very good. Meanwhile, the item reliability score was 0.89, meaning that the reliability of the resulting instrument was good [10]. From these results, the instrument will get a few different

results if the test is repeated over a long period and will produce consistent scores on each item. That way, this test instrument is reliable.

3. Difficulty Index

Difficulty index analysis was performed using the Output Table: Item Measure menu. The score that needs attention is located in the JMLE Measure column, showing the order of the difficulty level of the questions from the most difficult to the easiest based on the logit [10,31]. The results of the instrument difficulty index analysis can be seen in Figure 3.

Item STATISTICS: MEASURE ORDER

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	G
7	4	30	5.48	.65	.76	-.43	.40	-.50	.64	.50	89.7	89.6	3A	A
10	29	29	3.03	.40	.86	-.42	.86	-.39	.44	.58	75.0	72.2	4A	B
9	37	30	1.96	.38	1.31	1.17	1.34	1.25	.48	.54	62.1	70.4	3C	B
2	40	30	1.52	.38	1.53	2.05	1.48	1.80	.40	.51	62.1	67.7	1B	B
11	39	29	1.48	.39	.67	-1.56	.65	-1.55	.56	.52	82.1	67.4	4B	B
6	42	30	1.23	.38	.93	-.28	.99	.04	.40	.50	69.0	65.4	2C	B
5	51	30	-.21	.44	.97	-.03	.90	-.14	.41	.42	75.9	74.8	2B	B
1	28	30	-.30	.78	.53	-.80	.16	-.73	.54	.27	93.1	93.1	1A	A
4	28	30	-.30	.78	.53	-.80	.16	-.73	.54	.27	93.1	93.1	2A	A
8	28	30	-.30	.78	.53	-.80	.16	-.73	.54	.27	93.1	93.1	3B	A
3	54	30	-.86	.50	1.51	1.46	1.30	.67	.30	.38	82.8	82.6	1C	B
14	28	29	-1.09	1.06	1.14	.44	.66	.20	.17	.21	96.4	96.4	5B	A
15	54	29	-1.40	.59	.79	-.41	1.08	.36	.41	.35	92.9	87.4	5C	B
16	53	28	-1.73	.67	1.51	1.08	.65	-.11	.40	.34	88.9	89.4	6A	B
17	54	28	-2.25	.79	.96	.13	1.21	.57	.27	.29	92.6	92.6	6B	B
12	57	29	-3.12	1.06	1.13	.43	.66	.20	.17	.21	96.4	96.4	4C	B
13	57	29	-3.12	1.06	1.13	.43	.66	.20	.17	.21	96.4	96.4	5A	B
MEAN	40.2	29.4	.00	.65	.99	.10	.78	.02			84.8	84.0		
P.SD	14.3	.7	2.18	.24	.33	.92	.41	.78			11.5	11.3		

Figure 9. Difficulty index test results with Item Measure

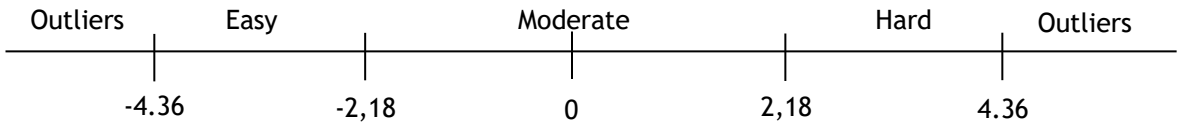


Figure 10. Difficulty level of questions

```
Person RAW SCORE-TO-MEASURE CORRELATION = .76
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .86  SEM = 1.40
STANDARDIZED (50 ITEM) RELIABILITY = .86
```

SUMMARY OF 17 MEASURED (NON-EXTREME) Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	40.2	29.4	.00	.65	.99	.10	.78	.02
SEM	3.6	.2	.55	.06	.08	.23	.10	.19
P.SD	14.3	.7	2.18	.24	.33	.92	.41	.78
S.SD	14.7	.7	2.25	.25	.34	.94	.42	.80
MAX.	57.0	30.0	5.48	1.06	1.53	2.05	1.48	1.80
MIN.	4.0	28.0	-3.12	.38	.53	-1.56	.16	-1.55

REAL RMSE	.74	TRUE SD	2.05	SEPARATION 2.79	Item	RELIABILITY	.89
MODEL RMSE	.70	TRUE SD	2.07	SEPARATION 2.97	Item	RELIABILITY	.90
S.E. OF Item MEAN = .55							

Figure 11. The results of the different power tests with Summary Statistics

Based on these results, this instrument's standard deviation (SD) was found to be 2.18. Therefore, a group of problem difficulties was obtained, as seen in Figure 9 above [27,32]. It can be concluded that items with a logit between -2.18 to 2.18 are questions with moderate difficulty. Items with a logit between -2.18 to -4.36 are easy questions. Items with a logit between 2.18 and 4.36 are difficult questions. Whereas items with a logit outside these values are outlier items (too easy/too difficult). Furthermore, data on item difficulty distribution can be seen in Table 3 [27,33].

From this analysis, it can be concluded that the instrument has a good difficulty level because it is dominated by items with moderate difficulty (not too difficult nor too easy).

Table 3. Item difficulty index

Kategori	Item Number	Logit	%
Outlier	3A	5.48	5,9%
Easy	4A	3.03	5,9%
Moderate	3C	1.96	70,6%
	1B	1.52	
	4B	1.48	
	2C	1.23	
	2B	-0.21	
	1A	-0.30	
	2A	-0.30	
	3B	-0.30	
	1C	-0.86	
	5B	-1.09	
	5C	-1.40	
	6A	-1.73	
Hard	6B	-2.25	17,6%
	4C	-3.12	
	5A	-3.12	

4. Discrimination Index

Differential power analysis was performed using the Output Table: Summary Statistics menu. The separation score obtained was 2.79. Thus, the differential power of the test instruments based on stratum calculations is H = {(4 x *Separation*) + 1}/3 = 4.05 (rounded up to 4.00). It proves that the items in the instrument have four different power types, namely very difficult, difficult, moderate, and easy questions [10].

So the discriminating power of this instrument is good. The results of the analysis obtained can be seen in Figure 11.

Reviewing the items fit

From the analysis above, it is evident that the validity, reliability, difficulty index, and discriminating power of all items in the instrument have met the fit criteria or are by the model. The validity and reliability of the instrument are good, the overall difficulty index of the items is balanced, and the differential power of the instrument is good.

Reviewing the Wright map

A review of the Wright map was carried out to see the distribution of the subject's ability to the difficulty level of the questions. The left axis shows the distribution of students' abilities, while the right axis shows the distribution of item difficulty. The subjects above have a high ability and understanding of the reaction rate material. The items above have a great difficulty level, and vice versa [34-36]. The results of Wright's map analysis of the instrument can be seen in Figure 12.

Based on a review of Wright's map, it was found that the subject with the highest ability was P18, who could answer all the questions correctly, even item number 3A, which was the most difficult. Meanwhile, the subjects with the lowest abilities were L11 and L20 because they could only answer items with moderate and easy difficulty levels.

Meanwhile, the test instrument found one outlier item, number 3A, because it is outside the T limit (with a logit of +5.48). The logit of question 3A exceeds the instrument's standard deviation limit, which is +4.36. It indicates that the question has a much higher difficulty level than the other items. So question number 3A must be revised or removed from the instrument. In addition, the other 16 questions have an even distribution of difficulty levels spread across -3 to +3 logits. There are four groups of problem distribution difficulties in this instrument, namely very difficult (3A), difficult (4A), moderate (3C, 1B, 4B, 2C, 1A, 2A, 2B, 3B, 1C, 5B, 5C, and 6A), and easy (6B, 4C, and 5A). Items 4C and 5A, although they are the easiest questions (with a logit of -3.12), do not require revision because both are still within the T limit. It can be concluded that the test instrument has good eligibility because moderate difficulty levels dominate the distribution of the items, so the questions in the instrument are relatively easy and not too easy [27].

Based on the results of the Wright map analysis, it was found that the student with the highest ability was P18. This student could answer all the questions correctly, even the most difficult items (3A). Meanwhile, students with the lowest abilities were L11 and L20 because they could only answer items with moderate and easy difficulty levels. From the distribution of abilities on the Wright map, most students in this study have moderate abilities.

The resulting test instrument can assess the overall level of student's understanding of the rates of reaction topic. A level 3 student has to successfully answer all sub-items in one question with a perfect score. Question 1, as exemplified above, tests students' understanding of concentration's effect on the reaction rate. This question's level of representation and its maximum scores.

Stated as follows: 1A (macroscopic, maximum score 1), 1B (symbolic, maximum score 2), and 1C (sub-microscopic, maximum score 2). Remember that the maximum score for this question is 1, 2, 2. Student L09 achieved the following scores: 1, 2, 2 (all sub-items answered correctly with perfect scores). Based on the analysis guidelines in Table 2, it can be concluded that L09 has a level 3 understanding of this matter. P30 students who scored 1, 0, and 2 (symbolic questions answered incorrectly) also had level 3 understanding because they could answer macroscopic-submicroscopic questions correctly. Whereas L15, whose scores of 1, 2, 1 (sub-microscopic questions did not achieve the

maximum correct score), had level 1 understanding because they could only answer macroscopic-symbolic questions correctly. In the first question (the effect of concentration on the rate of reaction), there were nine students (30%) who managed to answer all sub-items perfectly (macroscopic-submicroscopic-symbolic) correctly. There were 15 people (50%) who managed to answer the macroscopic-submicroscopic questions correctly. Based on the understanding level analysis guide compiled [21], both groups (80% of total subjects) have level 3 understanding. Meanwhile, only one student (3%) had level 2 understanding, for they could answer submicroscopic-symbolic questions correctly. Finally, three people (10%) answered macroscopic-symbolic questions correctly, and 2 (7%) answered only macroscopic questions. Therefore, there are a total of 17% of students have a level 1 understanding. From the analysis above, it can be concluded that the level of student's understanding of the effect of concentration on the rate of reaction is dominated by level 3 (80%), then level 1 (17%), and level 2 (3%) which indicates that the level students' understanding of that matters is good.

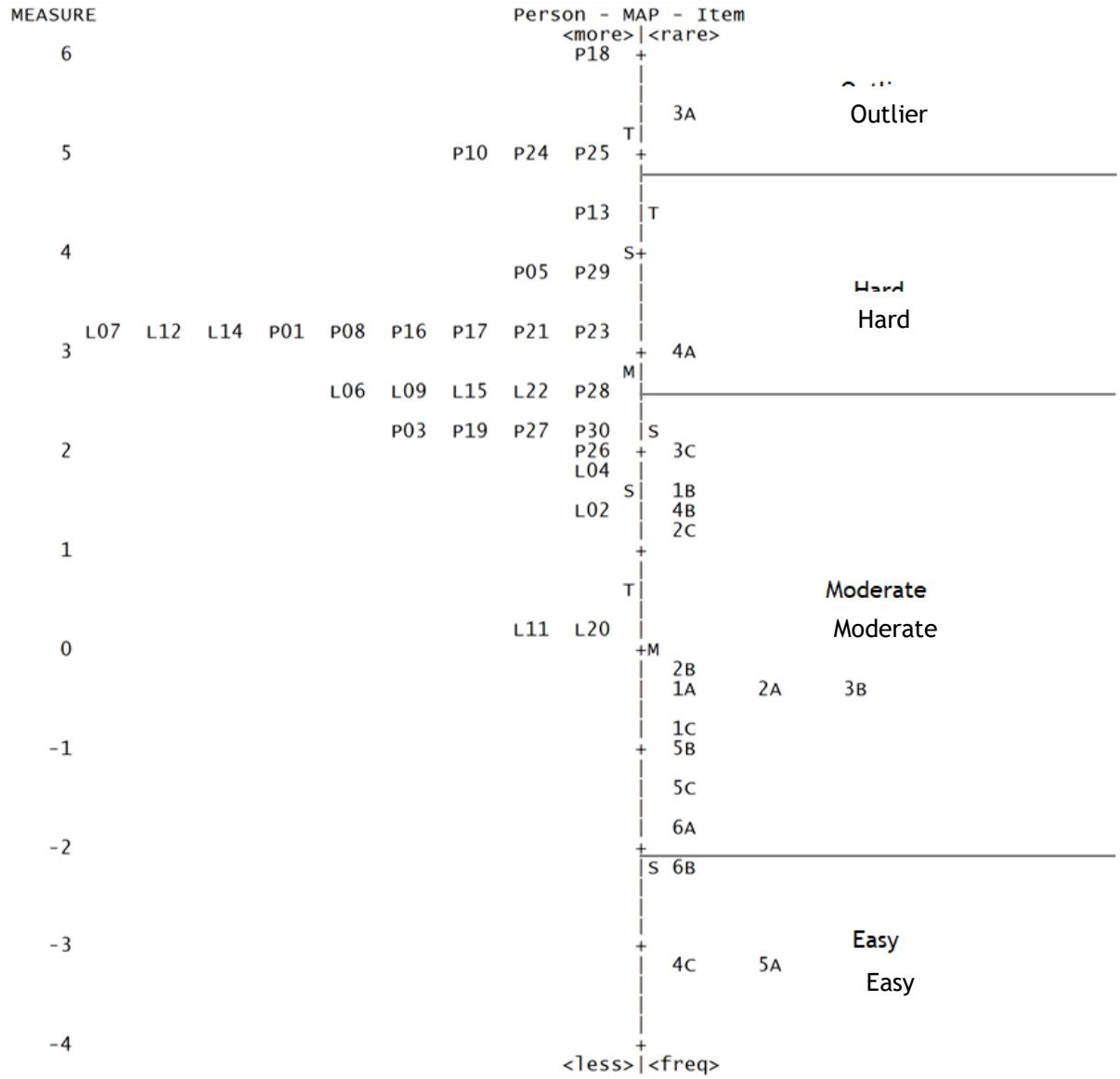


Figure 12. Wright Map

Repeating steps 4-7 until all items fit

Analysis of validity, reliability, difficulty index, and discriminatory power proves that all the items in this instrument are of good quality because they fit the expected criteria. Even so, from the results of a review of Wright's map, it was found that item number 3A was an outlier item. However, although this item required revision and retesting, a second trial was not conducted due to research time constraints.

Claiming the quality of the instrument

All items in the test instrument to test students' understanding of the macroscopic, sub-

microscopic, and symbolic levels of the reaction rate material proved high quality because they had been tested for validity, reliability, difficulty index, and discriminatory power.

Documenting the instrument

The documents provided in the results of this study contain important information related to the purpose of using test instruments, learning progression, item indicators, item items, and assessment rubrics, as well as guidelines for analyzing the level of students' understanding of the macroscopic, sub-microscopic, and symbolic levels of the material reaction rate.

CONCLUSION

The test instrument developed to test students' understanding of the macroscopic, sub-microscopic, and symbolic levels of the reaction rate material has been tested to be valid, reliable, and has a good index of difficulty and discriminating power. The exact agreements and expected agreements obtained from the validation by experts have a fragile score difference, namely 58.0% and 59.0%, which indicates that the model and estimates in the analysis are classified as fit. The strata value obtained from proving validity is 7.82, proving that the experts' assessment is classified as reliable. Empirical test analysis proves that all items meet the fit of the MNSQ, ZSTD, and Pt Measure scores, so this test instrument is valid. Instrument reliability proved to be good, with a score of 0.89. The instrument has four difficulty index groups dominated by the medium difficulty level (70.6%) and shows that the test instrument is relatively easy. The instrument has four strengths: very difficult, difficult, medium, and easy.

REFERENCES

[1] Fahriyah, A. R., & Wiyarsi, A. (2017). Multiple Representations Skill of High School Students on Reaction Rate Material. *The 2nd International Seminar on Chemical Education 2017, February 2018*, 192–210.

[2] Sukmawati, W. (2019). Analisis level makroskopis , mikroskopis dan simbolik mahasiswa dalam memahami elektrokimia Analysis of macroscopic , microscopic and symbolic levels of students in understanding electrochemistry. *Jurnal Inovasi Pendidikan IPA*, *5*(2), 195–204.

[3] Fahmi, T. N., & Fikroh, R. A. (2022). Pengembangan Modul Bermuatan Multirepresentasi pada Materi Hidrokarbon untuk SMA/MA. *Jurnal Inovasi Pendidikan Kimia*, *16*(1), 53–58.

[4] Ni Made Ary Suparwati. (2022). Analisis Reduksi Miskonsepsi Kimia dengan Pendekatan Multi Level Representasi: Systematic Literature Review. *Jurnal Pendidikan Mipa*, *12*(2), 341–348.

[5] Jespersen, N. D., Hyslop, A., & Brady, J. E. (2015). *Chemistry: The Molecular Nature of Matter*. John Wiley & Sons, Inc.

[6] Rahman, A. A., & Narsyah, C. E. (2019). *Evaluasi Pembelajaran*. Uwais Inspirasi Indonesia.

[7] Simamora, H., Hartono, H., & Effendi, E. (2021). Analisis Kualitas Butir Soal Buatan Guru Kimia Pada Tes Ujian Tengah Semester Ganjil Kelas XII MIPA. *Hydrogen: Jurnal Kependidikan Kimia*, *9*(1), 8.

[8] Rahayu, W., Putra, M. D. K., Rahmawati, Y., Hayat, B., & Koul, R. B. (2021). Validating an indonesian version of the what is happening in this class? (wihic) questionnaire using a multidimensional rasch model. *International Journal of Instruction*, *14*(2), 919–934.

[9] Nur, L., Nurani, L. A., Suryana, D., & Ahmad, A. (2020). Rasch model application on character development instrument for elementary school students. *International Journal of Learning, Teaching and Educational Research*, *19*(3), 437–459.

[10] Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. Trim Komunikata Publishing Home.

[11] Giguère, G., Brouillette-Alarie, S., & Bourassa, C. (2023). A Look at the Difficulty and Predictive Validity of LS/CMI Items With Rasch Modeling. *Criminal Justice and Behavior*, *50*(1), 118–138.

[12] Tesio, L., Caronni, A., Kumbhare, D., & Scarano, S. (2022). Interpreting results from Rasch analysis 1. The “most likely” measures coming from the model. *Disability and Rehabilitation*, *Submitted*(0), 1–13.

[13] Wei, S., Liu, X., Wang, Z., & Wang, X. (2012). Using rasch measurement to develop a computer modeling-based instrument to assess students’ conceptual understanding of matter. *Journal of Chemical Education*, *89*(3), 335–345.

[14] Boone, W. J., Yale, M. S., & Staver, J. R. (2014). Rasch Analysis in the Human Sciences. In *Rasch Analysis in the Human Sciences*.

[15] Indaryati, M Yusup, Nuraeni, Z., Novita Sari, & Meryansumayeka. (2022). Pelatihan dan Pendampingan Penyusunan IPK. *Jurnal Anugerah*, *3*(2), 77–85.

[16] Sunyono. (2015). *Model Multi Representasi*. Media Akademi.

[17] Pikoli, M., Sukertini, K., & Isa, I. (2022). Analisis Model Mental Siswa dalam Mentransformasikan Konsep Laju Reaksi Melalui Multipel Representasi. *Jambura Journal of Educational Chemistry*, *4*(1), 8–12.

https://doi.org/10.34312/jjec.v4i1.13515

[18] Jeklin, A. (2016). *Tes Uraian (Essay) Pada Evaluasi Hasil Pembelajaran Matematika*. *July*, 1–23.

[19] Jayanti, E. (2020). Instrumen Tes Higher Order Thinking Skill Pada Materi Kimia Sma. *Orbital: Jurnal Pendidikan Kimia*, *4*(2), 135–149.

[20] Sujak, K. B., Gnanamalar, E., & Daniel, S. (2018). Understanding of Macroscopic, Microscopic and Symbolic Representations Among Form Four Students in Solving Stoichiometric Problems. *MOJES: Malaysian Online Journal of Educational Sciences*, *5*(3), 83–96.

[21] Wang, Z., Chi, S., Luo, M., Yang, Y., & Huang, M. (2017). Development of an instrument to evaluate high school students' chemical symbol representation abilities. *Chemistry Education Research and Practice*, *18*(4), 875–892.

[22] Ihsan, H. (2016). Validitas Isi Alat Ukur Penelitian Konsep Dan Panduan Penilaiannya. *PEDAGOGIA Jurnal Ilmu Pendidikan*, *13*(2), 266.

[23] Puger, I. G. N. (2021). Pengujian Validitas Isi Tes Hasil Belajar Yang Dinilai Oleh Subject Matter Expert (Sme). *Daiwi Widya*, *8*(3), 1–15. https://doi.org/10.37637/dw.v8i3.819

[24] Oliva, J. M., & Blanco, Á. (2022). Rasch analysis and validity of the construct understanding of the nature of models in Spanish-speaking students. *European Journal of Science and Mathematics Education*, *11*(2), 344–359.

[25] Desnita, D., Yusmaita, E., Iswendi, I., & Iryani, I. (2021). Studi Tingkat Preferensi Panelis Terhadap Karakteristik Sensori Selai Kolang Kaling (Arenga Pinnata Fruits). *LOGISTA - Jurnal Ilmiah Pengabdian Kepada Masyarakat*, *5*(2), 75.

[26] Rizki, M., & Yusmaita, E. (2021). *Pengembangan Butir Soal Literasi Kimia pada Materi Ikatan Kimia Menggunakan Model Rasch*. *3*(2).

[27] Palimbong, J., Mujasam, M., & Allo, A. Y. T. (2019). Item Analysis Using Rasch Model in Semester Final Exam Evaluation Study Subject in Physics Class X TKJ SMK Negeri 2 Manokwari. *Kasuari: Physics Education Journal (KPEJ)*, *1*(1), 43–51.

[28] Van der Linden, W. J. (2011). Applying the Rasch Model. *International Journal of Testing*, *1*(3–4), 319–326.

[29] Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurements in the Human Sciences*. Routledge.

[30] Azizah, N., Suseno, M., & Hayat, B. (2022). Item analysis of the rasch model items in the final semester exam indonesian language lesson. *World Journal of English Language*, *12*(1), 15–26.

[31] Romdlon, N., Adi, M., Amaruddin, H., Maulana, H., Adi, M., & A, L. Q. (2022). *Validity and Reliability Analysis Using the Rasch Model to Measure the Quality of Mathematics Test Items of Vocational High Schools*. *11*(117).

[32] Ilfiandra, Nadhirah, N. A., Suryana, D., & binti Ahmad, A. (2022). Development and Validation Peaceful Classroom Scale: Rasch Model Analysis. *International Journal of Instruction*, *15*(4), 497–514.

[33] Purwana, U., Rusdiana, D., & Liliawati, W. (2020). Pengujian Kemampuan Menginterpretasikan Grafik Kinematika Calon Guru Fisika: the Polytomous Rasch Analysis. *ORBITA: Jurnal Kajian, Inovasi Dan Aplikasi Pendidikan Fisika*, *6*(2), 259.

[34] Yustiqvar, M., Hadisaputra, S., & Gunawan, G. (2019). Analisis penguasaan konsep siswa yang belajar kimia menggunakan multimedia interaktif berbasis green chemistry. *Jurnal Pijar Mipa*, *14*(3), 135-140.

[35] Irawan, J., Hadi, S., Zulandri, Z., Jamaluddin, J., Syukur, A., & Hadisaputra, S. (2021). Validating metacognitive awareness inventory (MAI) in chemistry learning for senior high school: A rasch model analysis. *Jurnal Pijar Mipa*, *16*(4), 442-448.

[36] Anggraini, N., & Yusmaita, E. (2021). Pemetaan Level Literasi Kimia Peserta Didik Kelas XI MIPA di SMAN 1 Lubuk Basung pada Materi Termokimia dengan Model Rasch. *Edukimia*, *3*(3), 147–154.