Journal of Educational Sciences
Journal homepage: https://jes.ejournal.unri.ac.id/index.php/JES

P-ISSN 2581-1657
E-ISSN 2581-2203

Design of Science Process Skills Instruments Reaction Rate Material: Rasch Model Approach

Difva Apriyani, Faizah Qurrata Aini*
Departemen of Chemistry, Faculty of Mathematics and Natural Sciences, Univeritas Negeri Padang, Padang 25131, Indonesia

ARTICLE INFO

Article history:

Received: 19 July 2024
Revised: 03 Jan 2025
Accepted: 04 Jan 2025
Published online: 24 Jan 2025

Keywords:

Assessment Instrument
Rasch Model
Reaction Rate
Science Process Skills

* *Corresponding author:*

E-mail: faizah_qurrata@fmipa.unp.ac.id

Article Doi:

Doi: https://doi.org/10.31258/jes.9.1.p.157-172

A B S T R A C T

The assessment process conducted by teachers tends to focus on understanding chemistry rather than the students' science process skills. This is because there are no available assessment instruments to measure students' science process skills, resulting in the chemistry learning outcomes in the independent curriculum not being fully achieved. This research aims to develop an assessment instrument that can measure students' science process skills on the topic of reaction rates. This instrument was developed using ten stages of the Rasch model and consists of 24 multiple-choice questions based on the aspects of science process skills in the independent curriculum. Field test results were analyzed using Ministep software, and it was found that all questions were declared valid because they met the MNSQ, ZSTD, and Pt Mean Corr criteria. The reliability value of 0.84 falls into the good category, and it also has good difficulty and discrimination indices according to the Rasch model. The results of the study indicate that the developed science process skills assessment instrument is suitable for use.

1. Introduction

Chemistry is the science that studies matter and the changes it undergoes (Chang & Overby, 2011). Based on Sudarmin (2015) Chemistry studies everything about substances, including their composition, structure and properties, changes, dynamics, and energetics, which require skill and reasoning. In chemistry learning, there are two interrelated aspects, namely chemistry as a product in the form of chemical knowledge in the form of facts, concepts, principles, laws, and theories, and chemistry as a scientific work process (Sudarmin, 2015).

Chemistry as a product and chemistry as a scientific work process are encompassed in the learning outcomes of chemistry in the independent curriculum, which is divided into two elements: understanding of chemistry and process skills. Understanding of chemistry includes all the material studied, while

process skills encompass the entire scientific process. The element of process skills is the method used to achieve an understanding of chemistry, so these two elements are presented as a cohesive whole (Kemendikbudristek, 2024). The independent curriculum encourages the development of broader skills for students, not just academic knowledge. In science education, this skill is known as science process skills (Sudarmin, 2015), scientific process skills are the embodiment of science as a process (Verawati et al., 2014).

Science process skills (SPS) are the basic abilities to acquire knowledge about scientific products (Suja, 2020). SPS emphasizes the ability of learners to independently discover knowledge based on learning experiences, laws, principles, and generalizations, thereby providing more opportunities to develop higher-order thinking skills (Sudarmin, 2015). Therefore, the assessment of SPS should not be overlooked. In reality, it was found that teachers placed more emphasis on assessing the knowledge or understanding aspects of students, without considering the students' SPS aspects (Ilmiah et al., 2020).

Based on the needs analysis conducted at three high schools in the city of Padang, it was found that chemistry teachers tend to focus only on assessing the aspect of chemistry understanding, while the SPS aspect is less attended to. The teacher believes that SPS can only be assessed when students conduct experiments in the laboratory. However, according to Tosun (2019), the assessment of SPS should involve the material within it. Furthermore, the teacher has not assessed all aspects of SPS in the independent curriculum; the teacher only observes students' SPS directly without using assessment instruments. As a result, teachers cannot observe students' SPS as a whole.

The science process skills of students can be measured using the SPS assessment instrument. The use of the SPS assessment instrument can provide teachers with information about the SPS that students have or have not mastered, allowing for the improvement of students' SPS (Mardliya et al., 2017). The assessment instrument created is an SPS assessment instrument based on the SPS aspects present in the independent curriculum The SPS assessment instrument must be tested for its validity and reliability, and have good difficulty and discrimination indices. The SPS assessment instrument that is created will then be analyzed using the Rasch model with the Ministep Windows program. This Rasch model has a special feature, namely the ability to produce a map that shows the distribution of students' abilities and the difficulty level of questions using the same measurement scale, as well as using a systematic response pattern to predict the presence of incomplete data or missing data (Sumintono & Widhiarso, 2015).

Several previous studies examining the SPS assessment instrument, namely Salmawati et al., (2023), have developed the SPS test instrument on thermochemistry material and assessed its validity, reliability, difficulty level, and discriminating power. Then the research by Ilmiah et al., (2020) developed the SPS test instrument on acid-base material and measured its validity, practicality, and reliability. Furthermore, the development of instruments for stoichiometry, solubility and solubility product, basic chemical laws, and acid-base titration

materials has also been carried out. Based on previous research on SPS assessment instruments, no SPS assessment instruments have been found for the reaction rate material in the independent curriculum. Therefore, this study aims to develop an assessment instrument that can measure students' SPS on reaction rate material using the Rasch model.

2. Methodology

The type of research is Research and Development (R&D) and uses the Rasch model development. This research applies the ten stages of the Rasch model as modified by Liu (2020), which include (1) determining the objectives and population, (2) determining the construct to be measured, (3) identifying the performance of the specified construct, (4) conducting a pilot test or field test, (5) conducting Rasch analysis, (6) reviewing item fit statistics, (7) reviewing the Wright map, (8) repeating steps 4-7 until all items fit, (9) establishing claims of validity and reliability of the instrument, and (10) developing documentation for the instrument. This research produced 24 items, and each item is aligned with the aspects of SPS found in the chemistry learning outcomes of the independent curriculum, with each item measuring only one aspect of SPS. The instrument trial was conducted at SMA Negeri 5 Padang with 69 phase F students who had studied the reaction rate material. The results of the instrument trial were then analyzed using the Rasch model to determine the quality of validity, reliability, difficulty index, and item discrimination of the test items on the instrument.

1) Validity

Validity in the Rasch model is known as item fit. Item fit explains whether the test item functions well or not in conducting the measurement. The Item Fit criteria can be seen in Table 1.

Table 1. Item Fit

Criteria	Interval	Explanation
Outfit Mean Square	0,5 < MNSQ < 1,5	Accepted
Outfit Z-Standard	-2,0 < ZSTD < +2,0	Accepted
Point Measure Correlation	0,4 < *Pt Measure Corr* < 0,85	Accepted

(Sumintono & Widhiarso, 2015)

2) Reliability

The reliability value of each test item can be seen in the Summary Statistics menu under the Item Reliability section, which is useful for determining the quality of each test item in the instrument. The Item Reliability Criteria can be seen in Table 2.

Table 2. Item Reliability

Item Reliability Value	Criteria
< 0,67	Weak
0,67 – 0,8	Enough
0,81 – 0,9	Good
0,91 – 0,94	Very Good
> 0,94	Special

(Sumintono & Widhiarso, 2015)

3) Difficulty Index

The difficulty index is analyzed using the Item Measure, which contains logit information for each question item. In addition, there is also a standard deviation value, which, when combined with the average logit value, allows the difficulty index to be categorized. The difficulty index category can be seen in Table 3.

Table 3. Category Difficulty Index Item

Logit Value	Category
Bigger than +1SD	Very Difficult
0,0 *logit* + 1SD	Difficult
0,0 *logit* – 1SD	Easy
Smaller than -1SD	Very Easy

(Sumintono & Widhiarso, 2015)

4) Discrimination Indices

The grouping of item discrimination can be seen from the separation value, which can identify groups of respondents and test items. To see the grouping of the difference power in detail, the following equation can be used:

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

(Sumintono & Widhiarso, 2015)

3. Results and Discussion

Determining Objectives and Population

The purpose of developing this instrument is to create a summative assessment tool that can measure students' SPS in the topic of reaction rates. This summative assessment is conducted after the learning ends, such as at the end of one topic of material (which can encompass one or more learning objectives) to determine the extent of students' achievement in one or more learning objectives (Anggraena et al., 2022). The population in this study is the phase F students of the XII grade at SMA Negeri 5 Padang. SMA Negeri 5 Padang was chosen because of the school's characteristics as one of the schools that have implemented the independent curriculum.

Determining the Construct to be Measured

The construct in the Rasch model refers to the attribute that is the focus of measurement or that the measuring instrument aims to measure (Wei et al., 2012). The construct measured in this study is the SPS based on its aspects in the chemistry learning outcomes of phase F of the independent curriculum, which consists of the aspects (1) observing, (2) questioning and predicting, (3) planning and conducting investigations, (4) processing, analyzing data and information, (5) evaluating and reflecting, and (6) communicating results (Kemendikbudristek, 2024).

Identifying the Performance of the Specified Construct

After the construct is determined, specific behaviors that describe each level of performance on the construct are needed in order to be identified. The specific behavior in this study is the learning objectives related to the content and context specific to the reaction rate material that has been taught in school. After the learning objectives are determined, they can be used to develop specific tests that are developed, including:

a. Determining the number of questions and the test format

The developed assessment instrument consists of 24 items. The format of the test used is a written test using paper and pen with multiple-choice questions. The use of paper and pen is considered more practical, easy, and affordable, and can help with the concentration and understanding of learners (Siegel, 2023). Multiple-choice tests are chosen because the assessment can be done easily, quickly, and objectively (Arifin, 2012).

b. Formulating question indicators

Creating question indicators aims to ensure that all questions align with the learning objectives. The preparation of question indicators is summarized in a question grid table. The outline serves as a guideline in creating questions for the test device (Arifin, 2012). In the question grid table, it includes learning objectives, indicators of SPS, question indicators, question formats, answer keys, and question numbers.

c. Creating an assessment rubric

The assessment rubric is a guide created to evaluate the quality of performance achievements and to focus attention on the competencies that students must master (Anggraena et al., 2022). In the assessment rubric created, the score given for a correct answer is 1 and the score for an incorrect answer is 0 because even though four or five answer choices are provided, there is only one correct answer (Sumintono & Widhiarso, 2015).

d. Testing the logical validity of the developed assessment instrument

Logical validity is conducted to review the instrument in terms of content, construct, language, and other additional rules. The validity of the instrument is conducted by experts who understand the developed instrument (Sugiyono, 2019). This assessment instrument was validated by three lecturers and two chemistry teachers using a validation sheet that included two options, namely "Appropriate" and "Inappropriate" on 12 predetermined criteria. The logical validity test data was then analyzed using Minifacet software, and the analysis results can be seen in Table 4.

Table 4. Summary of Expert Assessment Analysis on Question Items

Strata Value	Reliability	Exact Agreement	Expect Agreement
3,26	0,83	96,5%	96,8%

Table 4 contains a summary of the item quality, indicating that the assessment instrument is appropriate according to experts (validator). Based on Table 4, a strata value of 3.26 was obtained, indicating a reliable assessment by the experts. The expert reliability value obtained is 0.83, which falls into the good category. The exact agreement value is 96.5% and is not much different from the expected agreement value of 96.8%, which means that the expert assessment (validator) results are not significantly different from the estimated results, so the items on the instrument can be considered valid (Sick, 2013).

e. Improvement of assessment instruments

There are several suggestions provided by experts to serve as improvements for the assessment instrument. Based on those suggestions, improvements were made before being tested on the students.

Conducting a Trial

The assessment instrument that has been tested for validity is then trialed on a small group of students and accompanied by interviews. This trial was conducted with 9 students who had studied the reaction rate material with high, medium, and low abilities. The purpose of the variation in students' abilities is to determine whether all students' abilities can cover the test items or if most students with high abilities can answer the test items correctly while students with low abilities cannot answer those test items correctly (Ahmad, 2015). Then conduct interviews with each student regarding the questions they have completed. The purpose of conducting a trial with a small number of students and interviews is to see how students interpret the developed test items so that revisions can be made if necessary (Liu, 2020).

Conducting Rasch Analysis

At this stage, an analysis is conducted on the raw data obtained during the trial with a small number of participants (9 people). The data was analyzed using the

Rasch model with the help of the Ministep application. Here are the results of the analysis obtained:

a. Validity

Validity in the Rasch model is also known as item fit (item conformity). Item Fit explains whether the test item functions well in measuring (Sumintono & Widhiarso, 2015). The analysis of item validity is conducted using the *Output Tables menu:Item Fit Order*. Each question item must meet at least one of the three criteria in Table 1 to be considered valid. If only one of the three criteria in Table 1 is met, then the question item is retained and does not need to be changed (Sumintono & Widhiarso, 2015). The results of the validity analysis of the trial on 9 students can be seen in Figure 1.

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
20	5	9	-.18	.83	1.59	1.40	1.31	.62	A .34	.58	55.6	77.2	S20
22	3	9	1.28	.91	1.09	.33	1.57	.86	B .54	.63	88.9	81.5	S22
18	7	9	-1.59	.90	1.53	1.30	1.47	.76	C .15	.42	66.7	78.9	S18
13	2	9	2.19	1.01	1.51	.95	1.04	.47	D .41	.58	77.8	85.4	S13
23	4	9	.52	.85	1.09	.34	1.50	.86	E .53	.62	77.8	77.4	S23
16	3	9	1.28	.91	1.47	.97	1.35	.65	F .44	.63	66.7	81.5	S16
19	6	9	-.86	.84	1.29	.87	1.07	.43	G .39	.51	66.7	75.3	S19
7	2	9	2.19	1.01	1.26	.60	.87	.33	H .50	.58	77.8	85.4	S7
15	3	9	1.28	.91	1.24	.61	.90	.15	I .56	.63	66.7	81.5	S15
24	6	9	-.86	.84	1.22	.70	.95	.32	J .43	.51	66.7	75.3	S24
3	4	9	.52	.85	1.14	.46	.93	.12	K .58	.62	77.8	77.4	S3
12	7	9	-1.59	.90	1.12	.41	1.01	.48	k .35	.42	88.9	78.9	S12
2	7	9	-1.59	.90	.95	-.01	.57	.11	j .47	.42	66.7	78.9	S2
8	7	9	-1.59	.90	.80	-.41	.50	.04	i .53	.42	88.9	78.9	S8
9	7	9	-1.59	.90	.80	-.41	.50	.04	h .53	.42	88.9	78.9	S9
1	8	9	-2.57	1.12	.72	-.22	.33	-.19	g .45	.30	88.9	88.8	S1
14	8	9	-2.57	1.12	.72	-.22	.33	-.19	f .45	.30	88.9	88.8	S14
4	1	9	3.38	1.21	.59	-.53	.21	-.39	e .62	.45	88.9	88.6	S4
5	4	9	.52	.85	.59	-.96	.42	-.87	d .81	.62	77.8	77.4	S5
6	5	9	-.18	.83	.41	-1.84	.32	-.90	c .84	.58	100.0	77.2	S6
21	5	9	-.18	.83	.41	-1.84	.32	-.90	b .84	.58	100.0	77.2	S21
17	2	9	2.19	1.01	.33	-1.39	.20	-.52	a .84	.58	100.0	85.4	S17
MEAN	5.2	9.0	-.33	1.01	.99	.05	.80	.10			80.3	80.7	
P.SD	2.3	.0	1.91	.28	.38	.91	.44	.54			12.5	4.3	

Figure 1. Results of the Small-Scale Trial Validity

Based on Figure 1, items S12, S18, S19, and S20 tend to be unfit with a Pt Mean Corr score < 0,4, but all four are worth retaining because their MNSQ and ZSTD scores meet the criteria. Items S1, S4, S5, S6, S14, S17, S21, and S22 also tend to be unfit because they have MNSQ scores < 0,5 (for items S1, S4, S5, S6, S14, S17, and S21) and MNSQ > 1,5 (for item S22), but all eight items are also deemed worthy of retention because they have ZSTD and Pt Mean Corr scores that meet the criteria. Meanwhile, for items S2, S3, S7, S8, S9, S12, S13, S15, S16, S19, S23, and S24, they meet all three criteria, namely MNSQ, ZSTD, and Pt Mean Corr. The developed instrument consists of 24 items, but only 22 items are fit (appropriate), while the other 2 items, S10 and S11, are not detected, so items S10 and S11 need to be revised.

b. Reliability

Reliability explains the extent to which measurements conducted repeatedly can yield consistent results (Sumintono & Widhiarso, 2014). Reliability analysis is conducted using the *Output Tables menu: Summary Statistics*. Determining the reliability of the instrument is done by considering the item reliability score (red box). The reliability results of the items can be seen in Figure 2.

SUMMARY OF 24 MEASURED (EXTREME AND NON-EXTREME) Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	5.2	9.0	-.33	1.01				
SEM	.5	.0	.40	.06				
P.SD	2.3	.0	1.91	.28				
S.SD	2.4	.0	1.95	.29				
MAX.	9.0	9.0	3.38	1.88				
MIN.	1.0	9.0	-3.95	.83				

REAL RMSE 1.10 TRUE SD 1.56 SEPARATION 1.42 Item RELIABILITY .67
MODEL RMSE 1.05 TRUE SD 1.60 SEPARATION 1.53 Item RELIABILITY .70
S.E. OF Item MEAN = .40

Figure 2. Results of Small-Scale Trial Reliability

Based on the reliability results in Figure 2, it was obtained that the item reliability score is 0,67. In Table 2, it can be seen that if the score on the item reliability ranges from 0,67 to 0,80, then its reliability is sufficient or enough (Sumintono & Widhiarso, 2015). It can be concluded that the quality of the items in the instrument aspect of its reliability is sufficient.

c. Difficulty Index

The difficulty index analysis is conducted using the Output Tables: Item Measure menu. The results of the item 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
4	1	9	3.38	1.21	.59	-.53	.21	-.39	.62	.45	88.9	88.6	S4
7	2	9	2.19	1.01	1.26	.60	.87	.33	.50	.58	77.8	85.4	S7
13	2	9	2.19	1.01	1.51	.95	1.04	.47	.41	.58	77.8	85.4	S13
17	2	9	2.19	1.01	.33	-1.39	.20	-.52	.84	.58	100.0	85.4	S17
15	3	9	1.28	.91	1.24	.61	.90	.15	.56	.63	66.7	81.5	S15
16	3	9	1.28	.91	1.47	.97	1.35	.65	.44	.63	66.7	81.5	S16
22	3	9	1.28	.91	1.09	.33	1.57	.86	.54	.63	88.9	81.5	S22
3	4	9	.52	.85	1.14	.46	.93	.12	.58	.62	77.8	77.4	S3
5	4	9	.52	.85	.59	-.96	.42	-.87	.81	.62	77.8	77.4	S5
23	4	9	.52	.85	1.09	.34	1.50	.86	.53	.62	77.8	77.4	S23
6	5	9	-.18	.83	.41	-1.84	.32	-.90	.84	.58	100.0	77.2	S6
20	5	9	-.18	.83	1.59	1.40	1.31	.62	.34	.58	55.6	77.2	S20
21	5	9	-.18	.83	.41	-1.84	.32	-.90	.84	.58	100.0	77.2	S21
19	6	9	-.86	.84	1.29	.87	1.07	.43	.39	.51	66.7	75.3	S19
24	6	9	-.86	.84	1.22	.70	.95	.32	.43	.51	66.7	75.3	S24
2	7	9	-1.59	.90	.95	-.01	.57	.11	.47	.42	66.7	78.9	S2
8	7	9	-1.59	.90	.80	-.41	.50	.04	.53	.42	88.9	78.9	S8
9	7	9	-1.59	.90	.80	-.41	.50	.04	.53	.42	88.9	78.9	S9
12	7	9	-1.59	.90	1.12	.41	1.01	.48	.35	.42	88.9	78.9	S12
18	7	9	-1.59	.90	1.53	1.30	1.47	.76	.15	.42	66.7	78.9	S18
1	8	9	-2.57	1.12	.72	-.22	.33	-.19	.45	.30	88.9	88.8	S1
14	8	9	-2.57	1.12	.72	-.22	.33	-.19	.45	.30	88.9	88.8	S14
10	9	9	-3.95	1.88	MINIMUM MEASURE				.00	.00	100.0	100.0	S10
11	9	9	-3.95	1.88	MINIMUM MEASURE				.00	.00	100.0	100.0	S11
MEAN	5.2	9.0		1.01	.99	.05	.80	.10			80.3	80.7	
P.SD	2.3	.0	1.91	.28	.38	.91	.44	.54			12.5	4.3	

Figure 3. Results of the Difficulty Index Test in a Small-Scale Trial

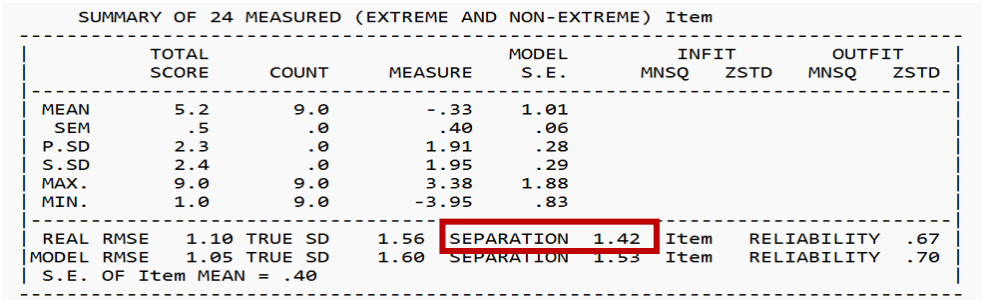
The difficulty index of the questions can be observed in the JMLE Measure column (red box), which contains logit values for each question item arranged from highest to lowest, indicating the order of question difficulty from hardest to easiest. The main benchmark in grouping the difficulty index is the standard deviation (SD) value (green box). The classification of the difficulty level of each item can be done by comparing the measure value of each item with the standard deviation (SD) value (SD) (Sabekti & Khoirunnisa, 2018). In Figure 3, a standard deviation value of 1,91 was obtained, allowing the difficulty level of each question item to be categorized. The data for grouping the test items based on their difficulty level can be seen in Table 5.

Table 5. Grouping of Test Items Based on Difficulty Level

Group	Logit	Question Item
Very Difficult	> 1,91	S4, S7, S13, and S17
Difficult	0,0 up to 1,91	S15, S16, S22, S3, S5, and S23
Easy	0,0 up to -1,91	S6, S20, S21, S19, S24, S2, S8, S9, S12, and S18
Very Easy	< -1,91	S1, S14, S10, and S11

d. Discrimination Indices

The analysis of item discrimination was conducted using the Output Tables: Summary Statistics menu. The results of the item discrimination analysis can be seen in Figure 4.



SUMMARY OF 24 MEASURED (EXTREME AND NON-EXTREME) Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	5.2	9.0	-.33	1.01				
SEM	.5	.0	.40	.06				
P.SD	2.3	.0	1.91	.28				
S.SD	2.4	.0	1.95	.29				
MAX.	9.0	9.0	3.38	1.88				
MIN.	1.0	9.0	-3.95	.83				

REAL RMSE	1.10	TRUE SD	1.56	SEPARATION	1.42	Item	RELIABILITY	.67
MODEL RMSE	1.05	TRUE SD	1.60	SEPARATION	1.53	Item	RELIABILITY	.70

S.E. OF Item MEAN = .40

Figure 4. Results of the Discrimination Indices in Small-Scale Trials

The initial step in determining the item discrimination index is to consider the value of separation, which is found to be 1,42. Next, the item discrimination test is conducted using the following stratum separation formula:

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

$$H = \frac{[(4\ x\ 1{,}42) + 1]}{3}$$

$$H = \frac{6{,}68}{3}$$

$$H = 2{,}2$$

The obtained stratum value is 2, which means that the instrument can distinguish between two groups of test items, namely difficult and easy (Sumintono & Widhiarso, 2015).

Review Item Fit Statistics and Revise If Necessary

Based on the analysis conducted on a small-scale trial involving a limited number of students, it was found that 22 items on the instrument were in accordance with the Rasch model, while 2 other items were invalid and needed to be revised. Analysis of reliability yielded an item reliability value of 0,67, which means the quality of the items is sufficient. Analysis of the difficulty index provided information on the variation of questions from very difficult, difficult, easy, to very easy. Meanwhile, the analysis of item discrimination a strata value of 2 was

obtained, which means that the instrument can distinguish between two groups of test items (difficult and easy).

Reviewing the Wright Map

The Wright map helps provide information about the distribution of students' abilities and the distribution of question difficulty levels on the same scale (Sumintono & Widhiarso, 2015). Based on the tests that have been conducted, a Wright map was obtained as shown in Figure 5.

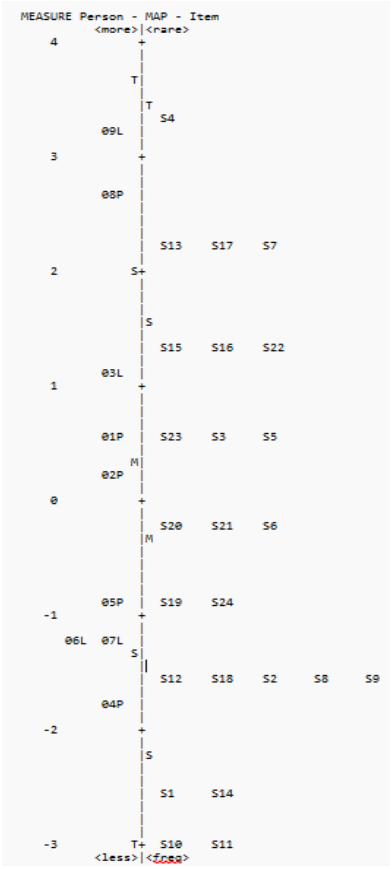


Figure 5. Results of the Wright Map Small-Scale Trial

The right section of the Wright map explains the distribution of item difficulty levels (Sumintono & Widhiarso, 2015). Item S4 is the question with the highest difficulty level (+3,38 logit) because it occupies the topmost position, which means that the chance of all students being able to answer this question correctly is very small. Meanwhile, items S10 and S11 are questions with the lowest logit values (-3,95 logit) and fall within the outlier range (beyond the T limit). The left part of the Wright map illustrates the distribution of students' abilities (Sumintono & Widhiarso, 2015). On the Wright map, it is seen that student 09L is at the very top, which means that the student has the highest ability, more than +3 logits, where they almost answered all the items correctly, even item S4, which is the item with the highest difficulty level.

Repeat Steps 4-7 Until All Items Fit

At this stage, steps 4-7 are repeated because the results of the data analysis obtained do not yet meet all the criteria of the Rasch modeling. There are two items that are invalid and outliers, so they need to be revised. The steps that need to be repeated are as follows:

a. Conducting Field Tests

After conducting trials with a small number of students, interviews, Rasch analysis, and revisions on the instrument items, the next step is to conduct field testing. The sample in this field test must represent the target population, which consists of 60 students with high, medium, and low abilities.

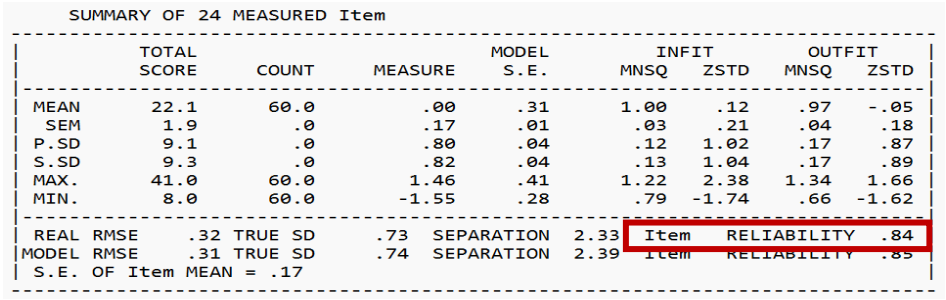
b. Rasch Analysis

At this stage, an analysis is conducted on the raw data obtained during the field test. This analysis was conducted to determine the validity (Figure 6), reliability (Figure 7), difficulty index (Figure 8), and discrimination power (Figure 9) of the questions on the assessment instrument.

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
9	15	60	.57	.32	1.22	1.29	1.34	1.38	A .09	.35	73.3	77.5	S9
19	26	60	-.38	.28	1.22	2.37	1.24	1.66	B .11	.35	50.0	65.4	S19
21	29	60	-.61	.28	1.20	2.38	1.19	1.33	C .13	.34	46.7	63.8	S21
20	21	60	.02	.29	1.16	1.35	1.16	.96	D .19	.35	61.7	69.7	S20
6	36	60	-1.14	.28	1.09	1.01	1.06	.41	E .23	.33	65.0	65.4	S6
1	41	60	-1.55	.29	1.06	.56	1.07	.37	F .24	.31	66.7	70.5	S1
2	33	60	-.91	.28	1.06	.69	1.07	.51	G .27	.33	61.7	63.8	S2
24	20	60	.11	.29	1.07	.59	1.04	.27	H .29	.35	66.7	70.9	S24
14	27	60	-.46	.28	1.04	.46	1.04	.37	I .30	.35	65.0	64.9	S14
17	8	60	1.46	.41	.98	.01	1.03	.20	J .34	.33	86.7	87.5	S17
22	12	60	.90	.35	1.03	.23	.99	.05	K .32	.34	78.3	81.7	S22
5	31	60	-.76	.28	1.01	.17	.98	-.07	L .33	.34	63.3	63.4	S5
11	26	60	-.38	.28	.95	-.48	1.01	.13	l .38	.35	70.0	65.4	S11
13	13	60	.79	.34	.89	-.51	1.00	.09	k .43	.35	81.7	80.3	S13
16	17	60	.38	.31	.99	-.04	.97	-.06	j .36	.35	73.3	74.9	S16
15	9	60	1.30	.39	.96	-.06	.72	-.71	i .42	.33	85.0	86.1	S15
4	29	60	-.61	.28	.95	-.57	.90	-.68	h .40	.34	60.0	63.8	S4
3	37	60	-1.22	.28	.93	-.76	.87	-.73	g .41	.32	70.0	66.2	S3
10	22	60	-.06	.29	.93	-.59	.86	-.88	f .44	.35	68.3	68.8	S10
18	14	60	.68	.33	.92	-.42	.89	-.37	e .43	.35	83.3	78.9	S18
12	18	60	.28	.30	.85	-1.08	.74	-1.42	d .53	.35	75.0	73.5	S12
23	17	60	.38	.31	.85	-1.01	.75	-1.28	c .53	.35	76.7	74.9	S23
7	11	60	1.02	.36	.79	-.97	.66	-1.13	b .57	.34	85.0	83.1	S7
8	19	60	.19	.30	.79	-1.74	.72	-1.62	a .58	.35	76.7	72.1	S8
MEAN	22.1	60.0	.00	.31	1.00	.12	.97	-.05			70.4	72.2	
P.SD	9.1	.0	.80	.04	.12	1.02	.17	.87			10.3	7.4	

Figure 6. Field Test Validity Results

Based on Figure 6, the items S9, S19, S21, S20, S6, S1, S2, S24, S14, S17, S22, S5, S11, and S16 tend to be unfit with a Pt Mean Corr score < 0,4, but these items are worth retaining because the MNSQ and ZSTD scores meet the criteria. Meanwhile, for the items S15, S4, S3, S10, S18, S12, S23, S7, and S8, they meet all the existing criteria, including MNSQ, ZSTD, and Pt Mean Corr. It can be concluded that all items on the instrument meet the valid or fit criteria, and no items need to be removed (Boone et al., 2014).



SUMMARY OF 24 MEASURED Item

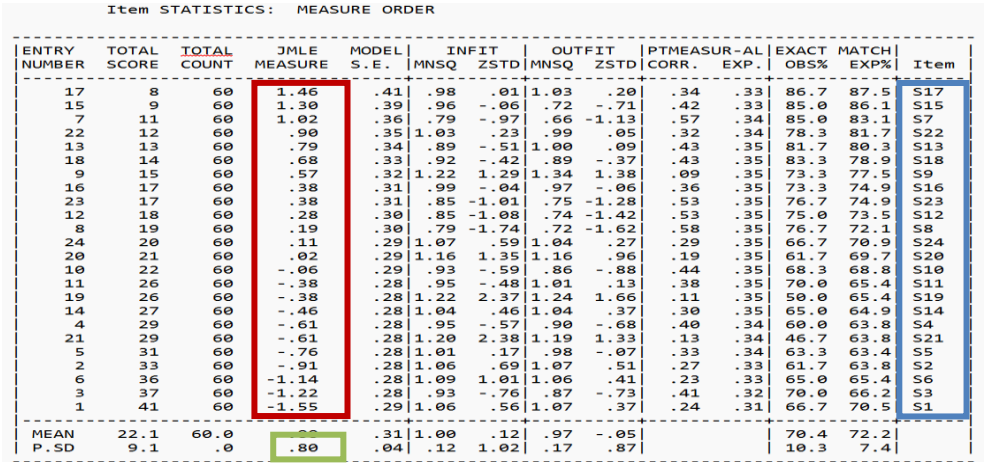
	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	22.1	60.0	.00	.31	1.00	.12	.97	-.05
SEM	1.9	.0	.17	.01	.03	.21	.04	.18
P.SD	9.1	.0	.80	.04	.12	1.02	.17	.87
S.SD	9.3	.0	.82	.04	.13	1.04	.17	.89
MAX.	41.0	60.0	1.46	.41	1.22	2.38	1.34	1.66
MIN.	8.0	60.0	-1.55	.28	.79	-1.74	.66	-1.62

REAL RMSE	.32	TRUE SD	.73	SEPARATION	2.33	Item RELIABILITY .84
MODEL RMSE	.31	TRUE SD	.74	SEPARATION	2.39	Item RELIABILITY .85

S.E. OF Item MEAN = .17

Figure 7. Field Test Reliability Results

Based on Figure 7, it was obtained that the item reliability score (red box) is 0,84. In Table 2, it is shown that if the score on the reliability item ranges from 0,8 to 0,9, then the reliability is good (Sumintono & Widhiarso, 2015). It can be concluded that the quality of the items in the instrument aspect of reliability is good.. This indicates that if it is repeated over a long period, the test results for each item will not differ significantly, thus the instrument produced is reliable (Sumintono & Widhiarso, 2015).



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
17	8	60	1.46	.41	.98	.01	1.03	.20	.34	.33	86.7	87.5	S17
15	9	60	1.30	.39	.96	-.06	.72	-.71	.42	.33	85.0	86.1	S15
7	11	60	1.02	.36	.79	-.97	.66	-1.13	.57	.34	85.0	83.1	S7
22	12	60	.90	.35	1.03	.23	.99	.05	.32	.34	78.3	81.7	S22
13	13	60	.79	.34	.89	-.51	1.00	.09	.43	.35	81.7	80.3	S13
18	14	60	.68	.33	.92	-.42	.89	-.37	.43	.35	83.3	78.9	S18
9	15	60	.57	.32	1.22	1.29	1.34	1.38	.09	.35	73.3	77.5	S9
16	17	60	.38	.31	.99	-.04	.97	-.06	.36	.35	73.3	74.9	S16
23	17	60	.38	.31	.85	-1.01	.75	-1.28	.53	.35	76.7	74.9	S23
12	18	60	.28	.30	.85	-1.08	.74	-1.42	.53	.35	75.0	73.5	S12
8	19	60	.19	.30	.79	-1.74	.72	-1.62	.58	.35	76.7	72.1	S8
24	20	60	.11	.29	1.07	.59	1.04	.27	.29	.35	66.7	70.9	S24
20	21	60	.02	.29	1.16	1.35	1.16	.96	.19	.35	61.7	69.7	S20
10	22	60	-.06	.29	.93	-.59	.86	-.88	.44	.35	68.3	68.8	S10
11	26	60	-.38	.28	.95	-.48	1.01	.13	.38	.35	70.0	65.4	S11
19	26	60	-.38	.28	1.22	2.37	1.24	1.66	.11	.35	50.0	65.4	S19
14	27	60	-.46	.28	1.04	.46	1.04	.37	.30	.35	65.0	64.9	S14
4	29	60	-.61	.28	.95	-.57	.90	-.68	.40	.34	60.0	63.8	S4
21	29	60	-.61	.28	1.20	2.38	1.19	1.33	.13	.34	46.7	63.8	S21
5	31	60	-.76	.28	1.01	.17	.98	-.07	.33	.34	63.3	63.4	S5
2	33	60	-.91	.28	1.06	.69	1.07	.51	.27	.33	61.7	63.8	S2
6	36	60	-1.14	.28	1.09	1.01	1.06	.41	.23	.33	65.0	65.4	S6
3	37	60	-1.22	.28	.93	-.76	.87	-.73	.41	.32	70.0	66.2	S3
1	41	60	-1.55	.29	1.06	.56	1.07	.37	.24	.31	66.7	70.5	S1
MEAN	22.1	60.0	.00	.31	1.00	.12	.97	-.05			70.4	72.2	
P.SD	9.1	.0	.80	.04	.12	1.02	.17	.87			10.3	7.4	

Figure 8. Results of the Difficulty Index Test in the Field Test

In Figure 8, a standard deviation value of 0,8 was obtained, allowing the difficulty level of each question item to be categorized. The data for grouping the test items based on their difficulty level can be seen in Table 6.

Table 6. Grouping of Test Items Based on Difficulty Level

Group	Logit	Question Item
Very Difficult	> 0,8	S17, S15, S7, and S22
Difficult	0,0 up to 0,8	S13, S18, S9, S16, S23, S12, S8, S24, and S20
Easy	0,0 up to - 0,8	S10, S11, S19, S14, S4, S21, and S5
Very Easy	< - 0,8	S2, S6, S3, and S1

Based on Table 6, it can be seen that the test items have a variation in difficulty levels ranging from very difficult, difficult, easy, to very easy, and are in accordance with the Rasch model. Based on Sudijono (2006) test items are considered good if their difficulty level is neither too hard nor too easy, in other

words, the degree of difficulty of the item is at a moderate level. Because in the Rasch model there is no group of moderate items, the difficult and easy items can be considered moderate items (Ahmad, 2015).

SUMMARY OF 24 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	22.1	60.0	.00	.31	1.00	.12	.97	-.05
SEM	1.9	.0	.17	.01	.03	.21	.04	.18
P.SD	9.1	.0	.80	.04	.12	1.02	.17	.87
S.SD	9.3	.0	.82	.04	.13	1.04	.17	.89
MAX.	41.0	60.0	1.46	.41	1.22	2.38	1.34	1.66
MIN.	8.0	60.0	-1.55	.28	.79	-1.74	.66	-1.62
REAL RMSE	.32	TRUE SD	.73	SEPARATION	2.33	Item	RELIABILITY	.84
MODEL RMSE	.31	TRUE SD	.74	SEPARATION	2.39	Item	RELIABILITY	.85
S.E. OF Item MEAN = .17								

Figure 9. Results of the Discrimination Indices in Field Trials

The separation value obtained in the field test is 2,33. Next, the item discrimination test is conducted using the following stratum separation formula:

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

$$H = \frac{[(4\ x\ 2{,}33) + 1]}{3}$$

$$H = \frac{10{,}32}{3}$$

$$H = 3{,}44$$

The obtained stratum value is 3, which means that the instrument can differentiate three groups of test items, namely difficult, moderate, and easy (Sumintono & Widhiarso, 2015).

c. Reviewing Item Fit Statistics

Based on the analysis that has been conducted, it was found that the validity, reliability, difficulty index, and discrimination power of all items in the instrument have met the fit criteria with the model. Analysis of validity shows that all items in the instrument are consistent with the model because they meet the MNSQ, ZSTD, and Pt Mean Corr criteria. Analysis of reliability yielded an item reliability value of 0,84, which means the quality of the items is good. Analysis of the difficulty index revealed that there are four variations of questions, namely very difficult, difficult, easy, and very easy. Meanwhile, for the item discrimination analysis, three discrimination items were obtained, namely difficult, moderate, and easy.

d. Reviewing the Wright Map

Based on Figure 10, it can be seen that the difficulty level of the test items has spread well compared to the small-scale trial. Item S17 is the question with the highest level of difficulty (+1,46 logit) because it occupies the topmost position,

which means that the chance of all students being able to answer this question correctly is very small.

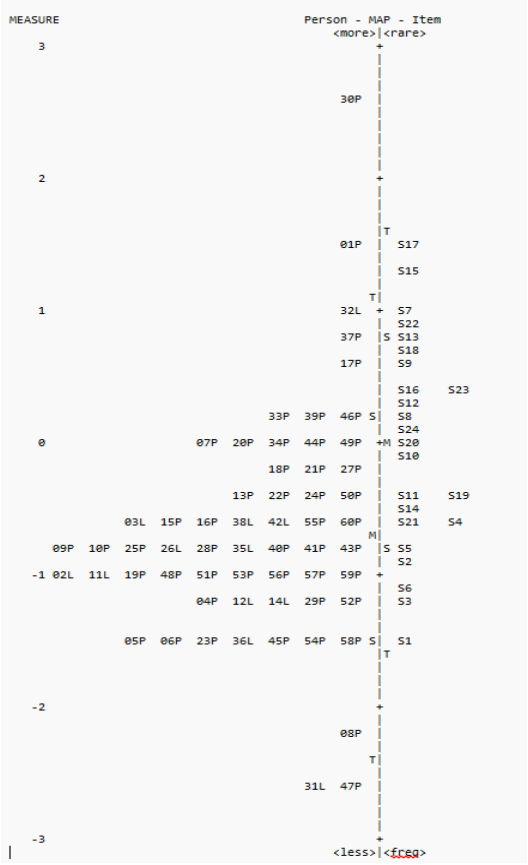


Figure 10. Field Test Wright Map Results

Meanwhile, item S1 is the question with the lowest logit value (-1,55) and is located at the very bottom, where almost all students can answer this question correctly. On the Wright map, it is seen that student 30P is at the very top, which means that this student has the highest ability compared to other students. Meanwhile, students 08P, 31L, and 47P are at the very bottom, which means that these students have lower abilities compared to other students.

Establishing Quality Claims for Questions

Based on data analysis from field tests, it was found that all items in the assessment instrument are proven to be of high quality because they have been tested for validity, reliability, difficulty index, and discrimination index, thus the assessment instrument can be used to measure SPS.

Developing Documentation for Instruments

Instrument documentation must be developed to facilitate users, whether teachers or students, in using the instrument (Liu, 2020). The provided documentation consists of the assessment instrument cover, the purpose of using the instrument,

usage instructions for teachers, a question grid, work instructions, a science process skills instrument sheet, and an assessment rubric.

4. Conclusion

Based on the research results, it can be concluded that the assessment instrument for science process skills on the reaction rate material that was developed has met the criteria of being valid, reliable, and having good difficulty and discrimination indices based on the Rasch model. Analysis of validity shows that all items have met at least one of the three validity criteria, namely MNSQ, ZSTD, and Pt Mean Corr. The reliability of the items falls within the good criteria, has four variations of questions, and has three item discrimination, so this assessment instrument can be used to measure students' science process skills.

References

Ahmad, N. (2015). *Buku Ajar Evaluasi Pembelajaran*. Yogyakarta: INTERPENA.

Anggraena, Y., Ginanto, D., & Felicia, N. (2022). *Panduan Pembelajaran dan Asesmen*. Badan Standar, Kurikulum, dan Asesmen Pendidikan Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia.

Arifin, Z. (2012). *Evaluasi Pembelajaran*. Jakarta: Direktorat Jenderal Pendidikan Islam Kementerian Agama RI.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. New York: Springer.

Chang, R., & Overby, J. (2011). *General Chemistry: The Essential Concepts, Sixth Edition*. New York: The McGraw-Hill.

Ilmiah, I., Anwar, M., & Herawati, N. (2020). Pengembangan Tes Keterampilan Proses Sains (KPS) pada Materi Asam Basa Kelas XI SMA/MA. *Chemistry Education Review, Prodi Pendidikan Kimia PPS UNM*, *4*(1), 64–70.

Kemendikbudristek. (2024). *Peraturan Menteri Pendidikan, Kebudayan, Riset, dan Teknologi Nomor 032/H/KR/2024* (Issue 021).

Liu, X. (2020). *Using and Developing Measurement Instruments in Science Education: A Rasch Modeling Approach*. USA: Information Age Publishing Inc.

Mardliya, S., Abdurachman, F., & Hartono. (2017). Pengembangan Instrumen Penilaian Keterampilan Proses Sains Dasar Mata Pelajaran Kimia Pada Kompetisi Dasar. *Jurnal Prosiding Seminar Nasional Pendidikan IPA*, *45*(1), 327–337.

Sabekti, A. W., & Khoirunnisa, F. (2018). Penggunaan Rasch Model untuk Mengembangkan Instrumen Pengukuran Kemampuan Berpikir Kritis Siswa Pada Topik Ikatan Kimia. *Jurnal Zarah*, *6*(2), 68–75.

Salmawati, L., Siswaningsih, W., Nahadi, & Rahmawati, T. (2023). Pengembangan Tes Keterampilan Proses Sains Kelas XI Pada Materi

Termokimia. *Jurnal Riset Dan Praktik Pendidikan Kimia*, *11*(2), 173–183.
Sick, J. (2013). Rasch Measurement in Language Education Rasch Measurement in Language Education Part 8: Rasch measurement and inter-rater reliability. *Shiken Research Bulletin*, *17*(2).
Siegel, J. (2023). Pen and paper or computerized notetaking? L2 English students' views and habits. *Computers and Education Open*, *4*(April 2022), 100120.
Sudarmin. (2015). *Model Pembelajaran Inovatif Kreatif [Model PAIKEM dalam Konteks Pembelajaran dan Penelitian Sains Bermuatan Karakter]*. Semarang: Swadaya Manunggal.
Sudijono, A. (2006). *Pengantar Evaluasi Pendidikan*. Jakarta: PT Raja Grafindo Persada.
Sugiyono. (2019). *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*. Bandung: Alfabeta.
Suja, I. W. (2020). *Keterampilan Proses Sains dan Instrumen Pengukurannya*. Depok: Rajawali Pers.
Sumintono, B., & Widhiarso, W. (2014). *Aplikasi Model Rasch untuk Penelitian Ilmu Ilmu Sosial*. Cimahi: Trim Komunikata.
Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Assesessment Pendidikan*. Cimahi: Trim Komunikata.
Tosun, C. (2019). Scientific Process Skills Test Development within the Topic "Matter and Its Nature" and the Predictive Effect of Different Variables on 7th and 8th Grade Students' Scientific Process Skill Levels. *Chemistry Education Research and Practice*, *20*(1), 160–174.
Verawati, N. N. S. V., Prayogi, S., & Asy'ari, M. (2014). Reviu Literatur Tentang Keterampilan Proses Sains. *Lensa: Jurnal Kependidikan Fisika*, *2*(1), 194.
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.

How to cite this article:

Apriyani, D., & Aini, F. Q. (2025). Design of Science Process Skills Instruments Reaction Rate Material: Rasch Model Approach. *Journal of Educational Sciences*, *9(1)*, 157-172.