HYDROGEN
JURNAL KEPENDIDIKAN KIMIA
https://e-journal.undikma.ac.id/index.php/hydrogen/index

December 2024. 12(6)
e-ISSN: 2656-3061
p-ISSN: 2338-6487
pp. 1441-1459

Rasch Model Analysis to Develop Assessment Instruments for Student's Science Process Skills on Chemical Bonding Material

Femi Oktavianti & Faizah Qurrata Aini*

Departement of Chemistry Education, Faculty of Mathematics and Natural Science, Padang State University, Jl. Prof. Dr. Hamka No.1, Padang, Indonesia 25173

* Corresponding Author e-mail: faizah_qurrata@fmipa.unp.ac.id

Article History
Received: 13-12-2024
Revised: 19-12-2024
Published: 31-12-2024

Keywords: assessment instruments, chemical bonds, rasch model, science process skills

Abstract

The assessment of student's science process skills (SPS) on chemical bonding material in schools does not yet meet the learning outcomes of the Merdeka curriculum, because the instruments used by teachers are more focused on knowledge aspects. The purpose of this research is to develop an assessment instrument for science process skills in the topic of chemical bonding. The type of this research is Research and Development (R&D) using the Rasch model modified by Liu. This research produced 18 questions tailored to the number of learning objectives in the chemical bonding material and the indicators of scientific process skills in the Merdeka curriculum, which consist of 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. The logical validity test was conducted by 5 validators consisting of 3 chemistry lecturers from FMIPA UNP and 2 chemistry teachers from State Senior High School (SMAN) 1 Banuhampu. The raw validation data were analyzed using Minifaced software, which showed that the data were valid according to the model. The raw trial data on 60 students from SMAN 1 Banuhampu were analyzed using Ministep, showing that all questions met the criteria of being valid, reliable, and having good difficulty and discrimination indices according to Rasch. These results indicate that the developed instrument is effective in measuring students' SPS on chemical bonding material and can be used to support learning based on a scientific approach.

How to Cite: Oktavianti, F., & Aini, F. (2024). Rasch Model Analysis to Develop Assessment Instruments for Student's Science Process Skills on Chemical Bonding Material. Hydrogen: Jurnal Kependidikan Kimia, 12(6), 1441-1459. doi:https://doi.org/10.33394/hjkk.v12i6.13941

https://doi.org/10.33394/hjkk.v12i6.13941

This is an open-access article under the CC-BY-SA License.

INTRODUCTION

Science education aims to engage students in the scientific process to understand concepts in depth. The scientific approach has become one of the relevant approaches to achieving this goal, as it emphasizes the scientific process such as data collection through observation or experimentation. This approach provides space for students to actively participate in learning (Indira, 2014). In the Merdeka curriculum, science process skills (SPS) become one of the elements of learning outcomes. This element is designed to train students in observing and communicating the results of chemistry learning (Permendikbud, 2024). This approach is believed to help students engage directly in the process of concept discovery, making learning more meaningful and capable of enhancing students' understanding (Ischak et al., 2020).

Chemistry learning often involves many abstract concepts related to the internal structure of matter, making chemistry a unique challenge for students to understand deeply. One of the

Copyright © HJKK, Authors

important topics in chemistry is chemical bonding, which serves as the foundation for understanding other concepts in chemistry learning. A good understanding of this material is essential to support further learning (Tsaparlis et al., 2018; Sitepu & Herlinawati, 2022). As one of the abstract concepts in chemistry, the teaching of chemical bonding requires the application of a scientific approach so that students can understand the concept in depth. The results of interviews with chemistry teachers at SMA Negeri 1 Banuhampu revealed that the discovery learning model has been used as part of the scientific approach in teaching this material. This indicates that students' science process skills have been trained in the learning process (Amelia & Syahmani, 2015; Hartini et al., 2017).

Although the scientific-based learning model has been implemented, the assessment of SPS elements in schools has not been carried out optimally. This constraint is due to the fact that the implementation of SPS assessment in the field faces several challenges. First, chemistry teachers more often focus on assessing knowledge aspects rather than process skills. This is influenced by the large number of students in one class, which makes individual observation and detailed evaluation difficult. In addition, the limited learning time makes it difficult for teachers to conduct SPS assessments, which require detailed observation, recording, and feedback. Then, the misunderstanding among teachers who often equate SPS with practical skills also becomes an obstacle, even though SPS encompasses a broader range of scientific activities, including theory and involving concepts (Hikmah et al., 2018;Tosun, 2019). This situation highlights the need for the development of assessment instruments that can comprehensively and validly measure students' scientific process skills (SPS).

Several previous studies have developed SPS assessment instruments on various chemistry topics such as Thermochemistry (Salmawati et al., 2023), Acids and Bases (Ilmiah et al., 2020), and Stoichiometry (Asmalia et al., 2015). Although most of these instruments have met the criteria for validity and reliability, there are limitations in the scope of the indicators used, such as a lack of reflection and evaluation. In addition, to date, there has been no development of SPS assessment instruments for the topic of chemical bonds. Therefore, there is a need for the development of more comprehensive instruments that align with the learning outcomes in the Merdeka curriculum.

In the development of this instrument, analysis using the Rasch model is the right choice because it can ensure validity, reliability, difficulty index, discrimination power, and item distribution based on student ability (Sumintono & Widhiarso, 2015). The Rasch model also provides objective data, thereby supporting the development of quality instruments. Based on the literature review, the application of the Rasch model in the development of instruments in chemistry subjects is still limited, particularly in the topic of chemical bonding. Therefore, this study aims to design a science process skills assessment instrument based on the Rasch model for the topic of chemical bonding, which is expected to support the implementation of a scientific approach and more effective learning evaluation in schools.

METHOD

This research uses the Research and Development (R&D) method with the Rasch development model. The development of the assessment instrument was conducted at the Department of Chemistry, FMIPA UNP. The trial of the developed instrument was conducted at State Senior High School (SMAN) 1 Banuhampu in the odd semester of the 2023/2024 academic year. In this study, the subjects involved were 5 validators consisting of 3 chemistry lecturers from FMIPA

UNP, 2 chemistry teachers, and 69 students from SMAN 1 Banuhampu. For data collection, this research utilized several instruments, such as: 1) Interview sheet: contains questions regarding the assessments conducted by teachers during lessons, teacher's views on SPS, and teacher's perceptions of the current level of students' SPS. 2) Logical validation test sheet: contains validation of material aspects, constructs, language, as well as suggestions for improvement and conclusions on the developed product. This validation uses the Guttman scale with "Yes" or "No" answers. 3) Instrument sheet for empirical validation test containing questions developed to measure student's SPS. This research procedure follows 10 steps in the development of instruments according to Liu, (2020) which consists of:

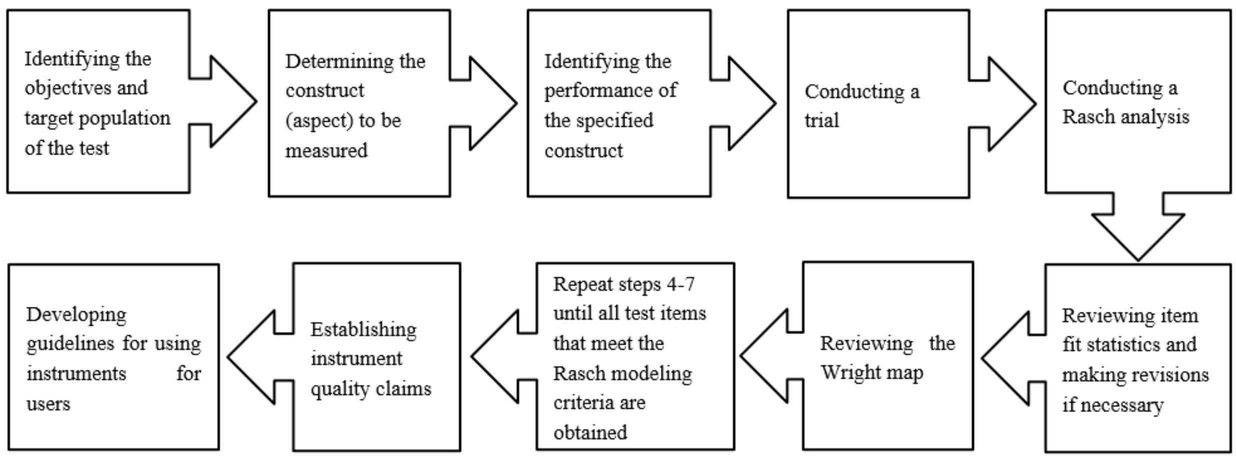

Figure 1. Research procedure

The data obtained were analyzed using the Ministep program. the data analysis techniques used to explain the research results are:

Table 1. Teknik analisis data

Aspect	Parámetro Rasch	Explanation
Validity	*Item fit* (*outfit* MNSQ, ZSTD, and *Pt Mean Corr*)	*Outfit* MNSQ 0,5 < MNSQ < 1,5 *Outfit* ZSTD -2,0 < ZSTD < +2,0 *Pt Mean Corr* 0,4 < Pt.M.C. <0.85
Reliability	*Summary statistisc (item reliability)*	<0,67 (weak) 0,67 – 0,80 (quite) 0,80 – 0,90 (good) 0,91- 0,94 (very good) 0,94 (special)
Difficulty index	*Item measure (logit scale)*	Greater than +1SD (very difficult) 0,0 logit + 1SD (difficult) 0,0 logit – 1SD (easy) Small from – 1SD (very easy)
Different power	$\mathrm{H} = \frac{[(4 \text{ x SEPARATION}) + 1]}{3}$	

RESULTS AND DISCUSSION

This research produced 18 questions used to measure students' science process skills (SPS) on the topic of chemical bonds. Each item is aligned with the construct and measures only one indicator of scientific process skills (SPS). The development of this instrument was carried out through the development stages based on Liu, (2020) as follows.

1. Identifying the Objectives and Target Population of The Test

The purpose of developing this instrument is to conduct a summative assessment on the elements of science process skills of Phase F students in 11th grade. Summative assessment is an evaluation conducted to determine the achievement of learning objectives. This assessment is conducted at the end of the material scope with two or more learning objectives. The purpose is to measure and determine students' learning achievements in one or more learning objectives over a specific period (Ginanto et al., 2024).

The population in this study is the 11th-grade students of Phase F at SMAN 1 Banuhampu. The selection of this school is based on its characteristics as one of the high schools that has already implemented the Merdeka curriculum and has a diverse student body, both in terms of academic abilities and social backgrounds (Kemendikbudristek, 2023). Then, the research sample was determined using the purposive sampling technique, which is a sample selection technique based on certain considerations (Sugiyono, 2019). In this study, the considerations in sample selection are the differences in students' abilities and the representation of the overall population, with the aim of ensuring that the research results can provide an accurate picture and represent all students' abilities.

2. Determining the Construct (aspect) To Be Measured

Construct refers to the attribute that is the focus of measurement from the instrument. Constructs can be abilities, skills, attitudes, or other attributes that cannot be measured directly, but rather through the subject's response to items (Wei et al., 2012). The construct in this study includes science process skills (SPS) in accordance with the SPS indicators in the Chemistry Phase F learning outcomes for 11th grade as outlined in BSKP Decision Number 032/H/KR/2024, 2024 as follows:

Table 2. Indicators of science process skills in the Merdeka curriculum

Indicator	Description
Observing	Students observe scientific phenomena and record their observations by paying attention to the details of the observed objects to generate questions that will be investigated.
Questioning and predicting	Students formulate scientific questions about the relationships between variables and hypotheses that can be investigated scientifically.
Planning and conducting an investigation	Students plan and select appropriate methods and control variables based on references to collect reliable data. Students choose and use tools and materials, including the appropriate use of digital technology, to systematically and accurately collect and record data.
Processing, analyzing data and information	Students interpret the information obtained honestly and responsibly. Students use various methods to analyze patterns and trends in the data. Students describe the relationships between variables and identify any inconsistencies that occur. Students use data and references to draw conclusions that are consistent with the investigation results.
Evaluation and reflection	Students identify sources of uncertainty and possible alternative explanations in order to evaluate conclusions and explain specifically to improve data quality. Students analyze the validity of information from primary and

Indicator	Description
	secondary sources and evaluate the approaches used to solve problems in the investigation.
Communicating the results	Students communicate the results of their investigation systematically and comprehensively, supported by scientific arguments and open to more relevant opinions.

3. Identifying the Performance of The Specified Construct

Identifying construct performance is an important step to ensure that the measurement instrument aligns with the characteristics being measured. After the construct is determined, specific behaviors that describe the subject's performance level on the construct must be identified. These specific behaviors are related to certain science content and context (Liu, 2020). In this study, the construct is based on students' science process skills in the topic of chemical bonding, so the behaviors representing this construct include the learning objectives in the topic of chemical bonding. Next, several specific stages were carried out to develop the test instrument, including:

1) Determining the Number and Format of the Questions

The questions are designed according to the learning objectives of the chemical bonding material and the indicators of science process skills (SPS) outlined in the Merdeka curriculum. The learning objectives for the chemical bonding material are threefold, including student's ability to distinguish between the formation processes of ionic and covalent bonds, students' ability to explain metallic bonds, and student's ability to relate the types of bonds to the properties of substances. Meanwhile, the SPS indicators consist of 6 indicators, so the number of questions in this study is 18 questions.

The format of the questions depends on the content and context to be measured, so the question format can be presented in various ways. The questions in this study were developed in the form of paper-and-pencil multiple-choice questions (Liu, 2020). This is based on the fact that its examination is more objective, with no elements of subjectivity influencing it, and it is easier and faster because an answer key can be used (Arikunto, 2009).

2) Development of Question Design

The development of question design begins with creating question indicators aimed at ensuring that all questions align with the learning objectives. Learning objectives are derived from the learning outcomes of the Merdeka curriculum by analyzing the competencies and scope of the material in the learning outcomes (Ginanto et al., 2024). The steps taken to derive learning objectives from learning outcomes are as follows:

a) Analysis of learning outcomes by identifying the competencies that must be achieved
b) Detailing competencies into indicators (based on the scope of the material)
c) Formulating specific learning objectives according to the details of the scope of the material

After the learning objectives are obtained, the next step is to formulate question indicators that include learning outcomes, learning objectives, indicators of science process skills, question indicators, and question items. Based on the indicators that have been compiled, 18 multiple-choice questions were produced, with each question representing one aspect of SPS.

3) Creating an Assessment Rubric

The assessment rubric is a guideline designed to evaluate and assess the quality of student competencies so that teachers can assist with the aim of improving student competencies. In addition, teachers can also use this assessment rubric to focus on the competencies that students need to master (Ginanto et al., 2024). The assessment rubric in this study includes learning objectives, SPS indicators, answer keys, and question weights.

4) Testing the Logical Validity of the Developed Instrument

Before the instrument is empirically tested on students, a logical validation test is first conducted, covering content, construct, language, and additional rules. The validity testing aims to ensure that the developed instrument aligns with the measurement objectives and learning goals, and is capable of measuring the expected construct (Arikunto, 2009). At this stage, it involves 5 validators (experts) consisting of 3 lecturers from the Chemistry Department of UNP and 2 chemistry teachers from SMAN 1 Banuhampu. This test is conducted by providing validation questionnaires and developed question indicators to the experts. The validation questionnaire consists of four assessed aspects, namely in terms of content, construct, language, and several additional rules. Then these four aspects are further developed into 12 criteria. The raw data obtained will be analyzed using Minifaced software with the following results:

```
+------------------------------------------------------------------------------------------------------------------+
| Total  Total  Obsvd  Fair(M)|   -    Model | Infit      Outfit    |Estim.| Correlation | Exact Agree. |          |
| Score  Count Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | Obs %  Exp % | N Pakar  |
|-----------------------------+--------------+----------------------+------+-------------+--------------+----------|
|  195    216     .90     .91 |  -2.30   .27 |  .69 -2.1   .36 -1.8 | 1.34 |   .58   .46 | 88.8   89.8  | 1 V1     |
|  210    216     .97     .98 |  -4.00   .44 |  .92  -.1  1.08   .4 | 1.01 |   .28   .27 | 93.5   93.8  | 2 V2     |
|  203    216     .94     .95 |  -3.01   .33 | 1.17   .8  1.96  1.4 |  .80 |   .31   .38 | 91.3   92.2  | 3 V3     |
|  210    216     .97     .98 |  -4.00   .44 | 1.20   .6  3.03  1.6 |  .82 |   .18   .27 | 92.8   93.8  | 4 V4     |
|  216    216    1.00    1.00 |( -7.18  1.84)|Minimum               |      |   .00   .00 | 94.7   94.7  | 5 V5     |
|-----------------------------+--------------+----------------------+------+-------------+--------------+----------|
|  206.8  216.0   .96     .97 |  -4.10   .67 |  .99  -.2  1.61   .4 |      |   .27       |              | Mean (Count: 5)     |
|    7.2     .0   .03     .03 |   1.67   .59 |  .21  1.2  1.00  1.4 |      |   .19       |              | S.D. (Population)   |
|    8.0     .0   .04     .04 |   1.86   .66 |  .24  1.4  1.15  1.6 |      |   .21       |              | S.D. (Sample)       |
+------------------------------------------------------------------------------------------------------------------+
   With extremes, Model, Populn: RMSE .89  Adj (True) S.D. 1.41  Separation 1.58  Strata 2.44  Reliability (not inter-rater) .71
   With extremes, Model, Sample: RMSE .89  Adj (True) S.D. 1.64  Separation 1.84  Strata 2.78  Reliability (not inter-rater) .77
Without extremes, Model, Populn: RMSE .38  Adj (True) S.D. .61  Separation 1.60  Strata 2.47  Reliability (not inter-rater) .72
Without extremes, Model, Sample: RMSE .38  Adj (True) S.D. .74  Separation 1.94  Strata 2.92  Reliability (not inter-rater) .79
With extremes, Model, Fixed (all same) chi-squared: 21.6  d.f.: 4  significance (probability): .00
With extremes, Model,  Random (normal) chi-squared: 2.8  d.f.: 3  significance (probability): .42
Inter-Rater agreement opportunities: 2160  Exact agreements: 1992 = 92.2%  Expected: 2005.8 = 92.9%
```

Figure 2. Results of the Expert Measurement Report analysis

Figure 2 shows a strata value of 2.92, meaning the validator's assessment is reliable, while the validator's reliability value is 0.79, which falls into the sufficient category. Then, the exact agreement value (validator agreement) is 92.2% and the expected agreement value (model estimate) is 92.9%, both of which have values that are not far apart, meaning the validator's assessment results are not significantly different from the model's estimate, thus the developed questions are categorized as appropriate (valid) (Boonee et al., 2014; Sick, 2013; Linacre, n.d.).This valid condition is considered to be met because the developed assessment instrument has been well-designed and in accordance with scientific theory or concepts (Arikunto, 2009). In this validation stage, the initial design of the test items receives suggestions from validators. Based on those suggestions, several improvements were made to the items to ensure the logic valid.

4. Conducting a Trial

The trial activities aim to obtain raw data that can be used to determine the validity, reliability, difficulty index, and discrimination power of the developed instrument. Before the trial was conducted, the students had already reviewed the topic of chemical bonds. The small-scale pilot test was conducted for 45 minutes (1 JP) on 9 students from 12th Grade class F 6 at SMAN 1 Banuhampu (3 students with high ability, 3 students with moderate ability, and 3 students with low ability) (Gall et al., 2003; Winarno ME, 2013). The purpose of differentiating the students' abilities is to represent the population in the instrument testing based on the Rasch model (Sugiyono, 2019). The purpose of this small-scale trial is to identify the feasibility of the test items, which is done to optimize the instrument before it is used in a large-scale trial (Gall et al., 2003; Kopp & Jones, 2020).

5. Conducting a Rasch Analysis

In this research, there is raw data processed using Ministep software with the following results:

a. Validity Test

The results of the validity analysis can be seen in the Item (column) menu: fit order. Each item is expected to meet the validity criteria, namely outfit MNSQ (0.5 – 1.5), outfit ZSTD (-2.0 – 2.0), and Pt. Mean Corr (0.4 – 0.85); the item must meet one of the established criteria to be considered fit (Sumintono & Widhiarso, 2015). The Outfit Mean Square value can assess the degree of randomness in the measurement system (Sumintono & Widhiarso, 2014), and the implications of the MNSQ score on measurements in the Rasch model are stated as follows.

Table 3. Implications of MNSQ values on measurement

MNSQ Value	Implications for measurement
>2,0	The questions are not easy to predict (underfit data), which can reduce the quality of the measurement system.
1,5 - 2,0	The instrument is not very good, but it doesn't lower the quality.
0,5 – 1,5	Good conditions for measurement
<0,5	The instrument is too easy to guess (data overfit), making it less productive for measurement, but it does not reduce quality.

Meanwhile, the Standardized fit statistic is a t-test to determine the instrument's fit with the model (Sumintono & Widhiarso, 2014), the implication of the ZSTD score on measurement in the Rasch model is stated as follows.

Table 4. Implications of ZSTD values on measurement

ZSTD Velue	Implications for measurement
≥3,0	Data is not expected to fit the model perfectly. However, in a large sample, the discrepancies may be smaller.
2,1 - 2,9	Data cannot be predicted.
-1,9 – 1,9	Data has a logical estimate.
≤-2,0	Data is too easy to predict.

Finally, the Pt. Mean Corr value can indicate misleading items (when subjects with low ability can answer correctly while subjects with high ability answer incorrectly). Values within the ideal range indicate that the instrument can effectively differentiate between students with high and low abilities, while values that are too high or too low indicate that the instrument is too easy (lacks variability) and does not support validity well (Boonee et al., 2014). The analysis results obtained are as shown in Figure 3.

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	7	9	-1.03	.94	1.58	1.68	3.70	1.63	A .14	.50	77.8	78.1	9
10	8	9	-2.06	1.13	1.49	.88	3.17	1.47	B-.01	.35	88.9	88.6	10
11	4	9	1.56	.96	1.74	1.23	2.31	1.40	C .45	.71	77.8	83.9	11
14	2	9	3.34	.98	1.50	1.17	.84	.35	D .42	.56	66.7	82.0	14
17	5	9	.67	.95	1.43	.83	1.19	.51	E .58	.70	77.8	84.6	17
12	8	9	-2.06	1.13	1.05	.29	.55	.08	F .33	.35	88.9	88.6	12
16	8	9	-2.06	1.13	1.05	.29	.55	.08	G .33	.35	88.9	88.6	16
2	8	9	-2.06	1.13	.88	.01	.40	-.09	H .40	.35	88.9	88.6	2
18	8	9	-2.06	1.13	.88	.01	.40	-.09	I .40	.35	88.9	88.6	18
3	6	9	-.19	.92	.77	-.40	.49	.01	h .71	.62	88.9	81.5	3
7	6	9	-.19	.92	.77	-.40	.49	.01	g .71	.62	88.9	81.5	7
4	7	9	-1.03	.94	.74	-.81	.39	-.10	f .60	.50	100.0	78.1	4
13	6	9	-.19	.92	.63	-.80	.39	-.12	e .75	.62	88.9	81.5	13
5	1	9	4.46	1.16	.53	-.69	.21	-.39	d .55	.41	88.9	88.2	5
6	4	9	1.56	.96	.33	-1.37	.22	-1.08	c .92	.71	100.0	83.9	6
8	5	9	.67	.95	.26	-1.58	.20	-1.00	b .92	.70	100.0	84.6	8
15	5	9	.67	.95	.26	-1.58	.20	-1.00	a .92	.70	100.0	84.6	15
MEAN	5.9	9.0	-.19	1.06	.94	-.07	.92	.10			88.2	84.4	
P.SD	2.1	.0	2.00	.22	.46	.98	1.05	.78			8.9	3.6	

Figure 3. Results of the validity test (small scale)

Based on Figure 3, there are 13 items that do not meet the MNSQ criteria (red box), namely S2, S3, S4, S5, S6, S7, S8, S13, S15, and S18 which have MNSQ values less than 0.5, and S9, S10, and S11 which have MNSQ values greater than 1.5. However, these items can be considered fit because they meet the ZSTD and Pt Mean Corr values. In addition, there are S6, S8, S15, S12, and S16 that do not meet the Pt Mean Corr value, but these two items are still considered fit because they meet the outfit MNSQ and ZSTD values. From Figure 12, it is known that all items in the fit category meet the ZSTD outfit value.

b. Reliability Test

Reliability analysis can be obtained using the Summary Statistics menu. The results of the analysis can be seen in Figure 4.

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	5.9	9.0	-.19	1.06				
SEM	.5	.0	.48	.05				
P.SD	2.1	.0	2.00	.22				
S.SD	2.2	.0	2.06	.22				
MAX.	9.0	9.0	4.46	1.87				
MIN.	1.0	9.0	-3.45	.92				

REAL RMSE 1.15 TRUE SD 1.63 SEPARATION 1.42 Item RELIABILITY .67
MODEL RMSE 1.08 TRUE SD 1.68 SEPARATION 1.55 Item RELIABILITY .71
S.E. OF Item MEAN = .48

Figure 4. Results of the reliability test (small scale)

The item reliability value of this instrument is 0.67, which means the reliability of the produced instrument is sufficient. Thus, the quality of the questions in the developed instrument can be trusted and reliable, because if repeated tests are conducted over a sufficiently long period, the results obtained will not differ significantly (Sumintono & Widhiarso, 2015; Walid & Ramli, 2015; Ngadi, 2023).

c. Difficulty Index Test

Difficulty index analysis is conducted using the Item Measure menu. The JMLE Measure column needs to be noted because it contains logits that indicate the order of question difficulty levels from the hardest to the easiest. The results of the difficulty index analysis of the developed instrument are shown in Figure 5.

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	1	9	4.46	1.16	.53	-.69	.21	-.39	.55	.41	88.9	88.2	5
14	2	9	3.34	.98	1.50	1.17	.84	.35	.42	.56	66.7	82.0	14
6	4	9	1.56	.96	.33	-1.37	.22	-1.08	.92	.71	100.0	83.9	6
11	4	9	1.56	.96	1.74	1.23	2.31	1.40	.45	.71	77.8	83.9	11
8	5	9	.67	.95	.26	-1.58	.20	-1.00	.92	.70	100.0	84.6	8
15	5	9	.67	.95	.26	-1.58	.20	-1.00	.92	.70	100.0	84.6	15
17	5	9	.67	.95	1.43	.83	1.19	.51	.58	.70	77.8	84.6	17
3	6	9	-.19	.92	.77	-.40	.49	.01	.71	.62	88.9	81.5	3
7	6	9	-.19	.92	.77	-.40	.49	.01	.71	.62	88.9	81.5	7
13	6	9	-.19	.92	.63	-.80	.39	-.12	.75	.62	88.9	81.5	13
4	7	9	-1.03	.94	.74	-.81	.39	-.10	.60	.50	100.0	78.1	4
9	7	9	-1.03	.94	1.58	1.68	3.70	1.63	.14	.50	77.8	78.1	9
2	8	9	-2.06	1.13	.88	.01	.40	-.09	.40	.35	88.9	88.6	2
10	8	9	-2.06	1.13	1.49	.88	3.17	1.47	-.01	.35	88.9	88.6	10
12	8	9	-2.06	1.13	1.05	.29	.55	.08	.33	.35	88.9	88.6	12
16	8	9	-2.06	1.13	1.05	.29	.55	.08	.33	.35	88.9	88.6	16
18	8	9	-2.06	1.13	.88	.01	.40	-.09	.40	.35	88.9	88.6	18
1	9	9	-3.45	1.87	MINIMUM MEASURE				.00	.00	100.0	100.0	1
MEAN	5.9	9.0	-.19	1.06	.94	-.07	.92	.10			88.2	84.4	
P.SD	2.1	.0	2.00	.22	.46	.98	1.05	.78			8.9	3.6	

Figure 5. Results of the difficulty index test (small scale)

In determining the difficulty index group of the instrument, the benchmark used is the standard deviation (SD) value. The SD value of this instrument is 2.00, resulting in the difficulty group of the questions as shown in Figure 6.

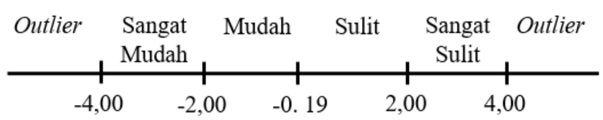

Figure 6. Level of difficulty of the question (small scale)

Based on Figure 6, the items with logit values in the range of 4.00 to 2.00 are classified as very difficult questions (S14). Items with logit values in the range of 0.19 to 2.00 are questions at the difficult level (S6, S8, S11, S15, and S17). Items with logit values in the range of -0.19 to -2.00 are considered easy questions (S3, S4, S7, S9, and S13). Meanwhile, questions with logit values in the range of -2.00 to -4.00 are classified as very easy questions (S1, S2, S10, S12, S16, and S18). Questions that fall outside the range of -4.00 to 4.00 are considered outlier questions (S5).

In addition, in Figure 6, there is a question (S1) with the minimum measure description. After further analysis, it turns out that all students were able to respond to the question correctly. Based on the results of the interviews with students in this trial, it was found that the answers could be guessed directly based on the provided images, so revisions need to be made.

d. Difference Test

The difference power analysis was conducted using the Summary Statistic menu, taking into account the Separation value. The results of the analysis obtained are as follows:

```
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE    S.E.      MNSQ   ZSTD  MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       5.9        9.0        -.19    1.06                               |
|  SEM        .5         .0         .48     .05                               |
| P.SD       2.1         .0        2.00     .22                               |
| S.SD       2.2         .0        2.06     .22                               |
| MAX.       9.0        9.0        4.46    1.87                               |
| MIN.       1.0        9.0       -3.45     .92                               |
|-----------------------------------------------------------------------------|
| REAL RMSE   1.15 TRUE SD    1.63  SEPARATION  1.42  Item   RELIABILITY  .67 |
|MODEL RMSE   1.08 TRUE SD    1.68  SEPARATION  1.55  Item   RELIABILITY  .71 |
| S.E. OF Item MEAN = .48                                                     |
```

Figure 7. Results of the small-scale differential power test

sBased on Figure 7, the Separation value of the developed instrument is 1.42, so the strata value (H) can be expressed as:

$$H = \frac{\{(4\ x\ Separation) + 1\}}{3}$$
$$= \frac{\{(4\ x\ 1{,}42) + 1\}}{3}$$
$$= \frac{\{5{,}68 + 1\}}{3}$$
$$= 2{,}23\ (2)$$

So, the items in this instrument have 2 levels of difficulty, namely difficult and easy (Sumintono & Widhiarso, 2015).

6. Reviewing Item Fit Statistics and Making Revisions if Necessary

The developed instrument is classified as valid even though there are several items that need to be revised, it has reliability in the sufficient category, the overall item difficulty index falls within the good criteria, and the discrimination index is in the sufficient category. Based on the objectives of the pilot test, it is necessary to revise item S1, which is not fit or does not conform to the model, and S5, as it is considered an outlier, before it is tested on a larger scale with students (Kopp & Jones, 2020).

Item S1 asks students to observe the image regarding the differences between the formation of ionic bonds and covalent bonds, then students are asked to determine the correct statement to express the differences between the two images presented. Based on interviews with 9 students who were the subjects of the study, it was found that item S1 became a minimum measure because the answer could be directly seen from the presented image and could be adjusted to the given options. So, revisions were made to the options to mislead the students, as shown in Figure 8.

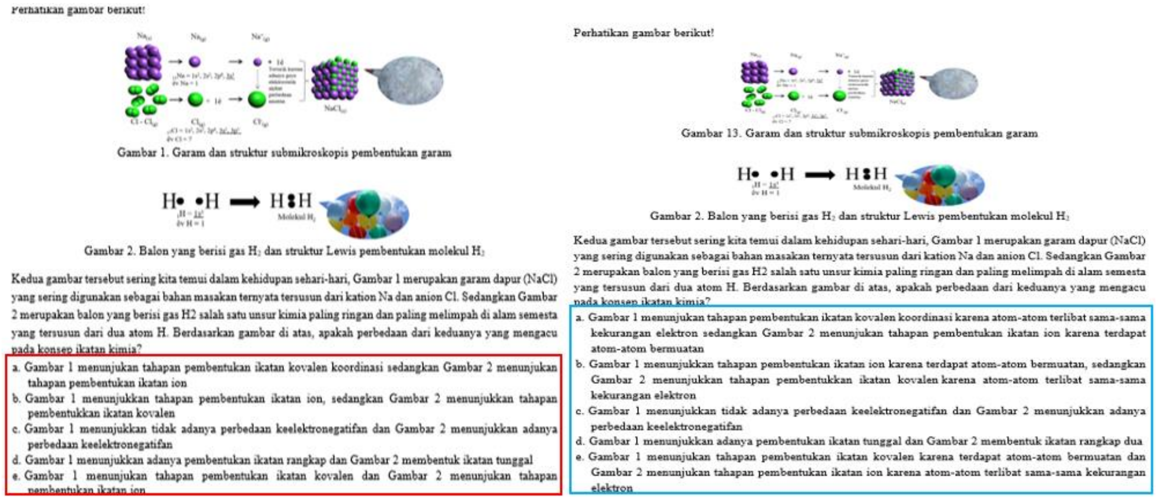

Figure 8. Item S1 before (red box) and after revision (blue box)

Meanwhile, item S5 is the question with the highest level of difficulty. In this question, the steps for forming ionic and covalent bonds are presented, and then the question asks students to identify the incorrect step in determining the bond formed between the two given elements. Based on interviews with 9 students who were the subjects of the study, it was found that item S5 was the most difficult because they determined the type of bond based on the given Lewis structure. They should pay attention to the details of both elements provided. So, the statement of the question was revised by adding the position/group of the element in the periodic table so that students can know whether the element is a metal or nonmetal and can determine the type of bond formed, as shown in Figure 9.

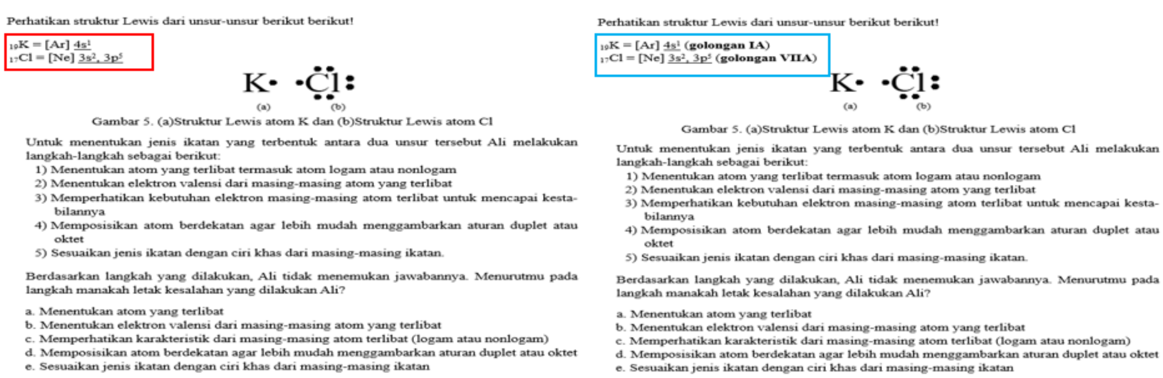

Figure 9. Item S5 before (red box) and after revision (blue box)

7. Reviewing the Wright Map

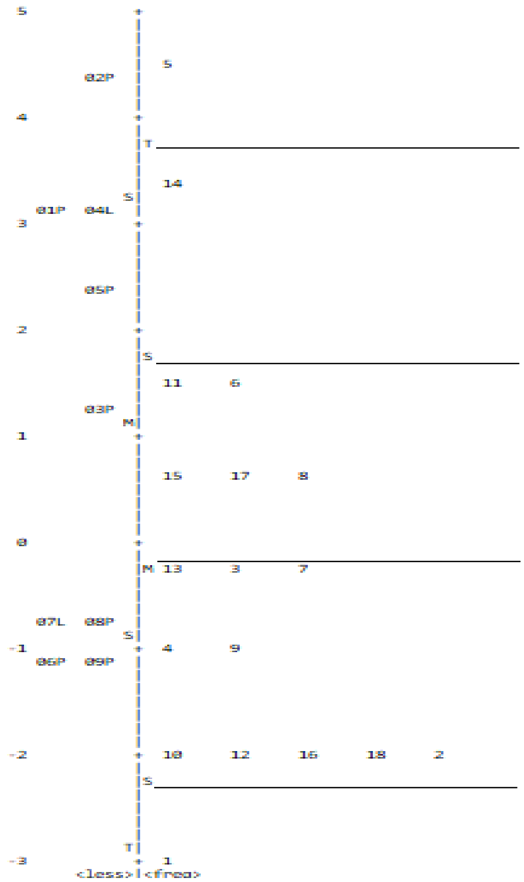

Figure 10. Map of Wright small-scale trial

The review of the Wright map aims to observe the distribution of student abilities against the difficulty level of the questions. The left side of the Wright map data shows the distribution of student abilities, while the right side shows the distribution of item difficulties. The results of Wright's analysis of the instrument can be seen in Figure 10.

The Wright map in this small-scale trial can be considered adequate, as the distribution of question difficulty levels sufficiently represents the students' abilities. However, in this map, there are still many items that overlap, meaning the difficulty levels of the questions in this instrument are the same (the logit values are the same). A good Wright map should have data distributed evenly without overlapping data (Bond & Fox, 2015). So, the developed instrument can be maintained on the grounds that each item represents different indicators and learning objectives, thus the items in this instrument need to be revised to improve the quality of the instrument and fill the gaps in the Wright map (Boonee et al., 2014).

8. Repeat Steps 4-7 Until All Test Items that Meet the Rasch Modeling Criteria are Obtained

The analysis of validity, reliability, difficulty index, and discrimination index in the small-scale trial indicates that there are several items that tend to be unfit but can still be retained with minor revisions before the instrument is tested on a larger scale. After the revisions were made, a large-scale trial was conducted for 60 minutes with 60 eleventh-grade students who chose the chemistry specialization at SMAN 1 Banuhampu. The size of the subjects is adjusted according to the sample size in Rasch modeling as explained by Linacre below:

Table 5. Sample size

Calibration of stable items in	Level of trust	Sample range
±1 *logit*	99%	27-61

(Sumintono & Widhiarso, 2014)

Based on the results of the large-scale trial, a Rasch analysis was conducted, with the results:

a) Validity Test

The results of the validity analysis can be seen in the Item menu (column): fit order. Each test item is expected to meet the validity criteria, namely outfit MNSQ (0.5 to 1.5), outfit ZSTD (-2.0 to +2.0), and Pt. Mean Corr (0.4 to 0.85), with at least one of these criteria being met by the test item (Sumintono & Widhiarso, 2015). The results of the analysis obtained are:

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
3	26	60	-.01	.28	1.34	3.18	1.49	3.22	A -.01	.38	46.7	66.7	P3
5	16	60	.85	.31	1.14	.96	1.35	1.43	B .15	.34	75.0	74.7	P5
13	26	60	-.01	.28	1.17	1.71	1.20	1.48	C .19	.38	60.0	66.7	P13
6	22	60	.31	.29	1.08	.77	1.17	1.06	D .26	.36	66.7	69.3	P6
17	11	60	1.39	.35	1.02	.17	1.16	.57	E .24	.30	83.3	81.7	P17
15	13	60	1.15	.33	1.02	.18	1.14	.57	F .27	.31	80.0	78.5	P15
9	34	60	-.64	.28	1.06	.61	1.11	.83	G .31	.38	61.7	66.5	P9
11	9	60	1.65	.38	1.03	.22	1.02	.19	H .23	.27	85.0	85.0	P11
8	23	60	.23	.29	1.02	.24	.96	-.22	I .36	.37	66.7	68.6	P8
2	24	60	.15	.28	.97	-.26	.96	-.25	i .41	.37	71.7	68.0	P2
1	46	60	-1.70	.33	.95	-.21	.86	-.46	h .40	.34	80.0	77.9	P1
4	41	60	-1.22	.30	.94	-.44	.94	-.25	g .43	.36	76.7	72.4	P4
10	22	60	.31	.29	.89	-1.00	.92	-.43	f .47	.36	73.3	69.3	P10
7	37	60	-.88	.29	.91	-.85	.85	-1.00	e .48	.38	73.3	68.7	P7
14	24	60	.15	.28	.90	-.97	.86	-.95	d .48	.37	71.7	68.0	P14
16	30	60	-.32	.28	.87	-1.41	.85	-1.27	c .52	.38	75.0	65.4	P16
18	25	60	.07	.28	.82	-1.87	.79	-1.62	b .57	.37	73.3	67.3	P18
12	44	60	-1.50	.31	.74	-1.87	.62	-1.88	a .64	.35	83.3	75.5	P12
MEAN	26.3	60.0	.00	.30	.99	-.05	1.01	.06			72.4	71.7	
P.SD	10.5	.0	.91	.03	.14	1.22	.21	1.23			9.3	5.7	

Figure 11. Results of the validity test (large scale)

Based on Figure 11, all items meet the MNSQ outfit value. There are 8 items that do not meet the Pt Mean Corr criteria (blue box), namely items S5, S6, S8, S9, S11, S13, S15, and S17 which have a Pt Mean Corr value less than 0.4. However, these items can be considered fit because they meet the MNSQ and ZSTD outfit values. Then, 1 item (S3) does not meet the ZSTD and Pt Mean Corr values, but item S3 is still considered fit because it has an MNSQ outfit value within the range of 0.5 to 1.5. Meanwhile, the other 9 items meet the three expected criteria. So, all the items in the developed instrument have proven to be valid, allowing these items to function normally in measuring students' science process skills (Sumintono & Widhiarso, 2015).

b) Reliability Test

Reliability analysis can be obtained using the Summary Statistic menu. The results of the analysis are shown in Figure 12. The item reliability value of this instrument is 0.88, meaning the reliability of the produced instrument is good. This indicates that if repeated testing is conducted over a sufficiently long period, the results obtained will be consistent. Thus, the instrument developed can be said to be reliable (Sumintono & Widhiarso, 2015).

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	26.3	60.0	.00	.30	.99	-.05	1.01	.06
SEM	2.5	.0	.22	.01	.03	.30	.05	.30
P.SD	10.5	.0	.91	.03	.14	1.22	.21	1.23
S.SD	10.8	.0	.94	.03	.14	1.26	.21	1.27
MAX.	46.0	60.0	1.65	.38	1.34	3.18	1.49	3.22
MIN.	9.0	60.0	-1.70	.28	.74	-1.87	.62	-1.88

REAL RMSE .31 TRUE SD .85 SEPARATION 2.76 Item RELIABILITY .88
MODEL RMSE .30 TRUE SD .86 SEPARATION 2.83 Item RELIABILITY .89
S.E. OF Item MEAN = .22

Figure 12. Results of the reliability test (large scale)

c) Difficulty Index Test

The analysis of the difficulty index is conducted using the Item Measure menu. The JMLE Measure column needs to be noted because it contains logits that indicate the order of question difficulty levels from the hardest to the easiest. The results of the difficulty index analysis of the developed instrument are as follows:

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
11	9	60	1.65	.38	1.03	.22	1.02	.19	.23	.27	85.0	85.0	P11
17	11	60	1.39	.35	1.02	.17	1.16	.57	.24	.30	83.3	81.7	P17
15	13	60	1.15	.33	1.02	.18	1.14	.57	.27	.31	80.0	78.5	P15
5	16	60	.85	.31	1.14	.96	1.35	1.43	.15	.34	75.0	74.7	P5
6	22	60	.31	.29	1.08	.77	1.17	1.06	.26	.36	66.7	69.3	P6
10	22	60	.31	.29	.89	-1.00	.92	-.43	.47	.36	73.3	69.3	P10
8	23	60	.23	.29	1.02	.24	.96	-.22	.36	.37	66.7	68.6	P8
2	24	60	.15	.28	.97	-.26	.96	-.25	.41	.37	71.7	68.0	P2
14	24	60	.15	.28	.90	-.97	.86	-.95	.48	.37	71.7	68.0	P14
18	25	60	.07	.28	.82	-1.87	.79	-1.62	.57	.37	73.3	67.3	P18
3	26	60	-.01	.28	1.34	3.18	1.49	3.22	-.01	.38	46.7	66.7	P3
13	26	60	-.01	.28	1.17	1.71	1.20	1.48	.19	.38	60.0	66.7	P13
16	30	60	-.32	.28	.87	-1.41	.85	-1.27	.52	.38	75.0	65.4	P16
9	34	60	-.64	.28	1.06	.61	1.11	.83	.31	.38	61.7	66.5	P9
7	37	60	-.88	.29	.91	-.85	.85	-1.00	.48	.38	73.3	68.7	P7
4	41	60	-1.22	.30	.94	-.44	.94	-.25	.43	.36	76.7	72.4	P4
12	44	60	-1.50	.31	.74	-1.87	.62	-1.88	.64	.35	83.3	75.5	P12
1	46	60	-1.70	.33	.95	-.21	.86	-.46	.40	.34	80.0	77.9	P1
MEAN	26.3	60.0	.00	.30	.99	-.05	1.01	.06			72.4	71.7	
P.SD	10.5	.0	.91	.03	.14	1.22	.21	1.23			9.3	5.7	

Figure 13. Results of the difficulty index test (large scale)

In determining the difficulty index group of the instrument, the benchmark used is the standard deviation (SD) value. The SD value of this instrument is 0.91, resulting in the following difficulty groupings for the questions:

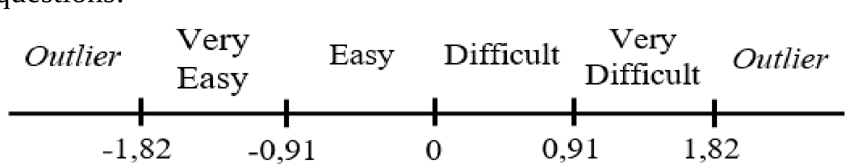

Figure 14. Level of difficulty of the question

Based on Figure 14. It can be determined that items with logit values between 0.91 and 1.82 are very difficult questions (S11, S15, and S17). Items with logit values in the range of 0 to 0.91 are considered difficult questions (S2, S5, S6, S8, S10, S14, and S8). Items with logit values in the range of 0 to -0.91 are considered easy questions (S3, S7, S9, S13, and S16). Meanwhile, questions with logit values in the range of -0.91 to -1.82 are classified as very easy questions (S1, S4, and

S12). In the Rasch mode, the difficulty index criteria consist of very difficult, difficult, easy, and very easy. Whereas very difficult and very easy questions are called outliers.

Based on Sudijono (2006), it is stated that an instrument item can be considered good if the question is neither too difficult nor too easy, thus the degree of item difficulty is moderate. So, questions that fall under the difficult and easy criteria in the Rasch analysis are considered medium difficulty questions. The distribution data of the question difficulty levels can be seen in Table 6.

Table 6. Distribution of the analysis results of question difficulty levels

Category	**Question number**	**Logit Velue**	**Percentage**
Difficult	11	1,65	16,7%
	15	1,39	
	17	1,15	
Currently	5	0,85	66,7%
	6	0,31	
	10	0,31	
	8	0,23	
	2	0,15	
	14	0,15	
	18	0,07	
	3	-0,01	
	13	-0,01	
	16	-0,32	
	9	-0,64	
	7	-0,88	
Easy	4	-1,22	16,7%
	12	-1,50	
	1	-1,70	

Based on Table 6, it is proven that the developed instrument has a good difficulty index, as the items overall are neither too difficult nor too easy (Sudijono, 2006). So, this instrument can accurately measure students' SPS and align with the measurement objectives, allowing it to be used to identify students' science process skills and comprehensively assess learning achievements.

d) Difference Test

The difference power analysis was conducted using the Summary Statistic menu while considering the Separation value. The analysis results obtained are as shown in Figure 13.

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	26.3	60.0	.00	.30	.99	-.05	1.01	.06
SEM	2.5	.0	.22	.01	.03	.30	.05	.30
P.SD	10.5	.0	.91	.03	.14	1.22	.21	1.23
S.SD	10.8	.0	.94	.03	.14	1.26	.21	1.27
MAX.	46.0	60.0	1.65	.38	1.34	3.18	1.49	3.22
MIN.	9.0	60.0	-1.70	.28	.74	-1.87	.62	-1.88

REAL RMSE .31 TRUE SD .85 SEPARATION 2.76 Item RELIABILITY .88
MODEL RMSE .30 TRUE SD .86 SEPARATION 2.83 Item RELIABILITY .89
S.E. OF Item MEAN = .22

Figure 13. Results of the large-scale differential power test

Based on Figure 13, the Separation value of this instrument is 2.80. The strata value (H) can be expressed as:

$$H = \frac{\{(4\ x\ Separation) + 1\}}{3}$$
$$= \frac{\{(4\ x\ 2{,}76) + 1\}}{3}$$
$$= \frac{\{11{,}04 + 1\}}{3}$$
$$= 4$$

So, it can be stated that the items in this instrument have 4 levels of difficulty, namely very difficult, difficult, easy, and very easy (Sumintono & Widhiarso, 2015). Based on the analysis that has been conducted, it has been proven that the validity, reliability, difficulty index, and discrimination index of all items in the instrument have met the fit criteria according to the model. The developed instrument has good validity and reliability, the overall item difficulty index meets the good criteria, and the discrimination power is excellent. The Wright map in this large-scale trial can be considered good, as the distribution of question difficulty levels adequately represents the students' abilities. So, the developed instrument can be maintained on the grounds that each item represents different indicators and learning objectives (Boonee et al., 2014).

Based on the review of the Wright map, it is observed that there is 1 student (48) with high ability. The logit score of 48 is outside the 2 standard deviation (T) limit, meaning student 48 has a significantly different high intelligence (outlier). Then, the student with low ability, namely student 22, has a logit value less than -2 standard deviations (T). Meanwhile, on the right side of the Wright map, the distribution of item difficulty levels is shown. Based on the review of the questions, P11 has the highest level of difficulty with a logit value still within 2 standard deviations.

The question with the lowest difficulty level is item P1, with a logit value also within -2 standard deviations. Meanwhile, the other items have a fairly even distribution of difficulty levels within the range of -2 logit to +2 logit, so the developed instrument can be considered good, as the distribution of item difficulty levels already represents the students' abilities. So, the developed instrument can be maintained on the grounds that each item represents different indicators and learning objectives, so the items in this instrument need to be revised to improve the quality of the instrument and fill the gaps on the Wright map (Boonee et al., 2014). The results of the Wright map analysis of the instrument can be seen in Figure 14.

9. Establishing Instrument Quality Claims

All items in the developed instrument aim to measure students' science process skills on the topic of chemical bonds, which have proven to be of high quality due to tested validity, reliability, difficulty index, and discrimination power. Validity and reliability tests are aimed at ensuring that the developed instrument can obtain data that reflects the measured ability and is accountable, so the assessment instrument should meet the requirements of content, construct, language, and have empirical validity evidence (Ngadi, 2023).

Overall, the developed instrument is valid and suitable for measurement. The analysis results also show an item reliability value of 0.88, meaning the quality of the developed test items is good. In addition, the resulting instrument has difficulty and discrimination indices that can differentiate students based on the level of question difficulty, such as very difficult, difficult, easy, and very easy (Sumintono & Widhiarso, 2015).

10. Developing Guidelines for Using Instruments for Users

The instrument usage guidelines provide the necessary information to assist users in applying the developed instrument (Sabekti & Khoirunnisa, 2018). This usage guide includes a cover (instrument identity), usage objectives, usage instructions for teachers, indicators of science process skills, question indicators, instructions for students, question items, and assessment rubrics for the developed instrument.

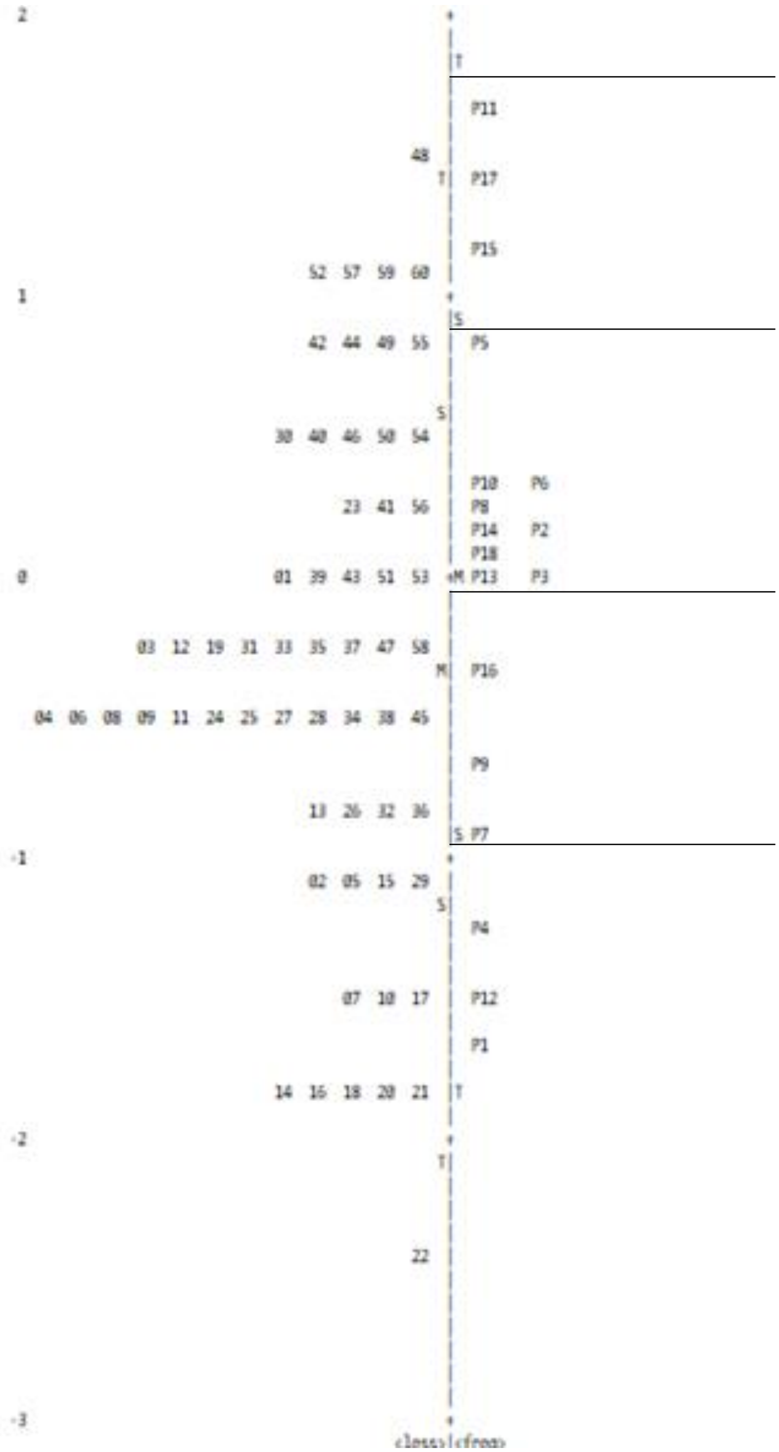

Figure 14. Map of Wright large-scale trial

CONCLUSION

Based on the results of the data analysis, all items meet the criteria of MNSQ, ZSTD, and Pt. Mean Corr, therefore the assessment instrument aimed at measuring students' science process skills on the topic of chemical bonds that has been developed has good validity, reliability, difficulty index, and discrimination according to the Rasch model. The developed instrument can be used by teachers to measure students' science process skills (SPS). With an objective and standardized assessment framework, this instrument can help teachers identify students' strengths and weaknesses in applying the scientific approach. In addition, the instrument design based on the Rasch model ensures valid and reliable assessment results, thereby supporting teachers in designing scientific-based learning strategies to enhance students' conceptual understanding and process skills. Not only that, this instrument can also serve as a reference for the development of similar instruments on other materials in various fields of science.

RECOMMENDATIONS

This study shows that the Rasch model-based science process skills (SPS) assessment instrument can be used to measure students' SPS on chemical bonding. Future research is expected to develop similar instruments for other chemistry topics. Furthermore, differences in instrument quality and analysis results can be influenced by factors such as student characteristics and the learning context in various schools, thus further research is needed to explore these factors.

ACKNOWLEDGEMENTS

We would like to thank those who have provided assistance, support, and contributions in the implementation of this research. Specifically, we extend our deepest gratitude to the supervising lecturer, who has provided direction, guidance, and valuable input in the execution of this research. We also extend our gratitude to the examiners, who provided constructive criticism and suggestions for the improvement of this research. Support from all parties is very meaningful for the success of this research.

BIBLIOGRAPHY

Amelia, D., & Syahmani. (2015). Meningkatkan Keterampilan Proses Sains dan Hasil Belajar Melalui Penerapan Pendekatan Scientific Materi Redoks Pada Siswa Kelas X MS 5 SMA Neneri 2 Banjarmasin. *QUANTUM, Jurnal Inovasi Pendidikan Sains*, *6*(2), 32–39.

Arikunto, S. (2009). DasarDasar Evaluasi Pendidikan (Edisi Revisi, Cet.10). Bumi Aksara.

Asmalia, I., Fadiawati, N., & Kadaritna, N. (2015). *Pengembangan Instrumen Asesmen Berbasis Keteramplan Proses Sains Pada Materi Stoikiometri*.

Bond, T. G. ., & Fox, C. M. . (2015). *Applying the Rasch model : fundamental measurement in the human sciences*. Routledge, Taylor and Francis Group.

Boonee·, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer.

Gall, M. D., Gall, J. P., & Borg, W. R. (2003). *Educational Research An Introduction* (A. E. Burvikovs, M. Kriener, & C. Tridente, Eds.; 7th Edition). Longman.

Ginanto, D., Kesuma, A. T., Anggraena, Y., & Setiyowati, D. (2024). *Panduan Pembelajaran dan Asesmen* (Edisi Revisi 2024). https://kurikulum.kemdikbud.go.id/file/1720050633_manage_file.pdf

Hartini, E. M., Kusasi, M., & Iriani, R. (2017). Meningkatkan Keterampilan Proses Sains dan Hasil Belajar Melalui Model Problem Solving Dengan Pendekatan Saintifik Pada Materi Hidrolisis Garam. *Journal of Chemistry And Education*, *1*(1), 37–45.

Hikmah, N., Yamtinah, S., Ashadi, & Indriyanti, N. Y. (2018). Chemistry teachers' understanding of science process skills in relation of science process skills assessment in chemistry learning. *Journal of Physics: Conference Series*, *1022*(1). https://doi.org/10.1088/1742-6596/1022/1/012038

Ilmiah, I., Anwar, M., & Herawati, N. (2020). Pengembangan Tes Keterampilan Proses Sains (KPS) pada Materi Asam Basa Kelas XI SMA/MA. *Chemistry Education Review*, *4*(1), 2597. https://doi.org/10.26858/cer.v4i1.13315

Indira, C. (2014). Best-Practices Pendekatan Saintifik Pada Pembelajaran Kimia Di SMA Negeri 4 Sampit. *Kaunia*, *X*(2), 141–151.

Ischak, N. I., Odja, E. A., Kilo, J. La, & Kilo, A. La. (2020). *Pengaruh Keterampilan Proses Sains Melalui Model Inkuiri Terbimbing terhadap Hasil Belajar Siswa pada* (Vol. 8, Issue 2). http://ojs.ikipmataram.ac.id/index.php/hydrogen

Kemendikbudristek. (2023). *Data Pokok Pendidikan*. Kemendikbudristek.

Keputusan BSKP Nomor 032/H/KR/2024, Pub. L. No. 032 (2024). https://guru.kemdikbud.go.id/dokumen/74r6Yln0zK?parentCategory=Implementasi%20Kurikulum%20Merdeka

Kopp, J. P., & Jones, A. T. (2020). Impact of Item Parameter Drift on Rasch Scale Stability in Small Samples over Multiple Administrations. *Applied Measurement in Education*, *33*(1), 24–33. https://doi.org/10.1080/08957347.2019.167430

Linacre, J. M. (n.d.). *Inter-rater and intra-rater Reliability*.

Liu, X. (2020). *Using and Developing Measurement Instruments in Science Education* (2nd ed.).

Ngadi. (2023). Analisis Model Rasch Untuk Mengukur Kompetensi Pengetahuan Siswa SMKN 1 KAalianget Pada Mata Pelajaran Perawatan Sistem Kelistrikan Sepeda Motor. *Jurnal Pendidikan Vokasi Otomotif*, *6*(1).

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., Rahmawati, T., Negeri, S., Wiranatakusumah, J., Bandung, K., Studi Pendidikan Kimia, P., & Pendidikan Matematika dan Ilmu Pengetahuan Alam, F. (2023). Pengembangan Tes Keterampilan Proses Sains Kelas XI Pada Materi Termokimia Development of Class XI Science Process Skills Test on Thermochemical Material. *Riset Dan Praktik Pendidikan Kimia*, *11*(2).

Sick, J. (2013). *Rasch Measurement in Language Education Rasch Measurement in Language Education Part 8: Rasch measurement and inter-rater reliability*.

Sitepu, D. S. B., & Herlinawati. (2022). Pengembangan Media Pembelajaran Berbasis Web Google Sites Pada Materi Ikatan Ion dan Kovalen Untuk SMA Kelas X. *Jurnal Ilmiah Pendidikan*, *1*(5)

Sudijono, A. (2006). *Pengantar Evaluasi Pendidikan*. PT RajaGrafindo Persada.

Sugiyono. (2019). *Metode Penelitian Kuantutatif, Kualitatif, dan R&D* (Sutopo, Ed.; Kedua). Alfabeta.

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi model rasch : untuk penelitian ilmu-ilmu sosial*. Trim Komunikata Publishing House.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch Pada Assessmen Pendidikan* (B. Trim, Ed.; I). 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. https://doi.org/10.1039/c8rp00071a

Tsaparlis, G., Pappa, E. T., & Byers, B. (2018). Teaching and Learning Chemical Bonding: Research-Based Evidence for Misconceptions and Conceptual Difficulties Experienced By Students In Upper Secondary Schools And The Effect of An Enriched Text. *The Royal Society of Chemistry*.

Walid, A., & Ramli, M. (2015). *Penyusunan Instrumen Tes High Order Thinking Skills Seminar Nasional XII Pendidikan Biologi FKIP UNS*.

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. https://doi.org/10.1021/ed100852t

Winarno ME. (2013). Metodologi Penelitian dalam Pendidikan Jasmani. In *ResearchGate* (Cetakan II). UM Press.