JOURNAL OF EDUCATIONAL SCIENCES

FKIP - University of Riau Publisher

Journal of Educational Sciences

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

P-ISSN 2581-1657

E-ISSN 2581-2203

# Development of Scientific Process Skill Assessment Instrument on Stoichiometry Using Rasch Model in the Merdeka Curriculum

**Nadia Pertiwi, Faizah Qurrata Aini***

*FMIPA, Padang State University, Padang, 25131, Indonesia*

**ARTICLE INFO**

*Article history*:

Received: 18 Oct 2024
Revised: 03 Jan 2025
Accepted: 04 Jan 2025
Published online: 24 Jan 2025

*Keywords:*

Test Instrument
Stoichiometry
Science Process Skill
Rasch Model

* *Corresponding author:*

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

*Article Doi:*

Doi: https://doi.org/10.31258/jes.9.1.p.241-255

**ABSTRACT**

The achievement of learning outcomes in the Merdeka Curriculum with respect to the assessment of scientific process skills remains unobservable. Teachers frequently conduct assessments that focus primarily on students' knowledge, while assessments targeting scientific process skills are rarely implemented. This is concerning because scientific process skills is part of the learning outcomes expected in the Merdeka Curriculum. Furthermore, there is a lack of assessment instruments in the form of tests that could assist teachers in measuring students' scientific process skills. This study used a research and development method using the Rasch model. The assessment instrument consisted of 24 multiple choice questions. The study involved three chemistry lecturers from the FMIPA at Universitas Negeri Padang, two chemistry teachers, and students from SMA Negeri 5 Padang as research subjects. Field test results were analyzed using Ministep software, which demonstrated that all items met the criteria for validity (based on outfit MNSQ, ZSTD, and Pt Mean Corr), reliability, varied difficulty levels (ranging from very difficult to very easy), and good item discrimination indices.

## 1. Introduction

The merdeka curriculum is a curriculum that focuses on the talent and interest of students where they are expected to choose any subject they like to explore the potential in themselves. The purpose of this curriculum is to create meaningful and effective learning that will foster the creation, taste, and spirit of students with lifelong Pancasila character. One of the subject matters contained in the merdeka curriculum is chemistry. There are two interconnected aspects in chemistry. First, chemistry as a product is a collection of facts, concepts, principles, laws, and theories. Second, chemistry as a process is a scientific work process where students discover and develop their own knowledge. In chemistry learning, process and product should be the top priority. Good product results come from a good learning process (Trianto, 2009).

Learner learning outcomes are measured from three component this is knowledge, attitude and skills. The skills component relates to the capacity to take action after a particular learning experience. The ability to directly address scientific problems with reasoning is called science process skills (Chetachukwu & Joshua, 2022). These skills are the basic ability to gain knowledge about the products of science (Suja, 2020). Teachers argue that these skills must be possessed by learners because it can help them find information independently through direct observation that allows them to understand and deepen what they learn. This will make students more active and participate during learning (Hamadi et al., 2018; Mahmudah, 2017). Science process skills are also listed in the demands of learning outcomes in the merdeka curriculum in addition to chemical understanding.

Based on the result of interviews with chemistry teachers of UNP Laboratory Development High School, SMAN 2 Padang, and SMAN 5 Padang, it can be concluded that the assessment is often done on the aspect of chemical understanding only while the assessment of process skill is rarely done. Whereas science process skills are also a demand of learning outcomes in the merdekacurriculum. Assessment of process skills aspects is only done when students do practicum, which is taken from the results of reports and the way students work in the field. In fact, the assessment of science process skills is not necessarily done when observing students in practicum but can also use question instruments (Amali & Firman, 2024). When assessing science process skills, the material must be included. Meanwhile, teachers still generally assess science process skills through observation during learning without involving the material being taught (Tosun, 2019). In addition, there is a lack of time to assess students' process skills and there are no test instruments that can help teachers assess students. As a result, the assessment of science process skills has not seen its achievement in the merdeka curriculum. For this reason, an instrument is needed that can help teachers assess students' process skills.

Assessment instruments can be in the form of written, oral and observation tests (Ramadhani et al., 2015). Although it can be done with several tests, written tests in the form of multiple choices are the right choice because multiple choice tests include objective tests (Basuki et al., 2019). The scoring system will produce the same score. In addition, it can make it easier for teachers to assess, save time, and minimize the use of tools and materials (Ramadhani et al., 2015). In using the assessment instrument, an analysis is needed to show the quality of the test to be used. One way to analyze assessment instruments is to use the Rasch model.

The Rasch model can provide an overview of the ability of students and the difficulty level of each item. This will make it easier for teachers to identify students' abilities. The five measurement principles used in the Rasch model are able to produce linear measures with equal distances, predict data loss, produce more accurate assessments, identify model inaccuracies, and produce measuring instruments that do not depend on the parameters being studied (Sumintono & Widhiarso, 2015)

Previously, a science process skills assessment instrument on stoichiometry material was developed by Asmalia et al. (2015). There are four indicators that are assessed, namely observing, inferring, predicting, and communicating. Furthermore, a science process skills assessment instrument has also been developed by Kristiyanto et al. (2019) in the form of a computerized test on stoichiometry material that refers to the science process skills contained in PISA. This test assesses five indicators, namely observation, interpreting data, predicting, applying concepts, and making conclusions. From the research that has been done, there is no instrument for assessing science process skills in stoichiometry material that refers to process skills indicators in the merdeka curriculum, so this can be a starting point for stating that this test needs to be developed in order to match the indicators of science process skills in the merdeka curriculum. Stoichiometry explains the calculation of the amount of reactants and products in a chemical reaction (Mahaffy, et al., 2022). This material underlies other materials such as thermochemistry, equilibrium, and acid-base. Based on these problems, this study developed a test instrument to measure science process skills on stoichiometry material using the Rasch model.

## 2. Methodology

Research and Development is a type of research conducted using the Rash model. The stages in this study were modified from the 10 stages of instrument development by Liu (2020) including (1) determining the population and objectives; (2) determining the construct to be measured (3) identifying the performance of the specified construct; (4) conducting trials or field tests; (5) conducting Rasch analysis; (6) reviewing item fit statistics; (7) looking at Wright's map; (8) If the item does not fit the Rasch model, repeat steps 4-7; (9) determining item quality claims; and (10) developing instrument documentation.

This research involved three chemistry lecturers from FMIPA Padang State University, two chemistry teachers, and SMAN 5 Padang students as research subjects. The data analysis technique of the research results was analyzed using minifac and ministep software in terms of validity, reliability, difficulty index, and differentiability. The validity test can be analyzed on the item fit menu by paying attention to three criteria, namely the outfit MNSQ, ZSTD, and Pt Mean Corr values. Reliability is analyzed on the summary statistic menu by paying attention to the item reliability value. The difficulty index is analyzed on the item measure menu. And the question's differential power is analyzed on the summary statistic menu by looking at the separation value. Another equation that can be used to see the grouping of items more thoroughly is called stratum separation using the following formula:

$$H = \frac{(4\ x\ separation) + 1}{3}$$

## 3. Results and Discussion

### *Setting the Popolation and Objectives*

The purpose of this research is to develop an assessment instrument to measure student's science process skills in stoichiometry material. The assessment instrument developed is a summative assessment because it aims to measure the achievement of learning objectives for stoichiometry material based on indicators of science process skills. Summative assessment is carried out at the end of the learning process after completing one subject matter with the aim of knowing the learning outcomes of students based on the learning objectives that have been set (Ratnawulan & Rusdiana, 2014). This study involved phase F students from class XI of SMA Negeri 5 Padang.

### *Determining the Construct to e Measured*

The construct to be measured in this study is science process skill taken from the learning outcomes in the merdeka curriculum. One of the requirements for constructs to be measured is linearity (Liu, 2020). The linearity of science process skills refers to the indicators of science process skills contained in the learning outcomes of the merdeka curriculum, namely, (1) observing; (2) questioning and predicting; (3) planning and choosing methods; (4) processing, analyzing data and information; (5) evaluating and reflecting; and (6) communicating results.

### *Identifying the Performance the Defined Constructs*

After the construct is determined, it is necessary to have specific behaviors that describe each level of performance on the construct so that it can be identified (Liu, 2020). The specific behavior in question is the learning objectives on stoichiometry material taught in class XI phase F. If the learning objectives have been identified, they can be used to develop test specifications. There are several steps that need to be taken to develop test specifications. First, determine the number of questions and test format. The number of items is designed based on the learning objectives that have been set. There are four learning objectives that have been identified, each with six indicators of science process skills. So the number of questions developed in this study was 24 multiple choice questions. The multiple choice test form is the right choice because multiple choice tests include objective tests (Basuki et al., 2019). The scoring system will produce the same score. In addition, it can make it easier for teachers to assess, save time, and minimize the use of tools and materials (Ramadhani et al., 2015)

Second, developing question indicators that include learning objectives, science process skills indicators, question indicators, question items, and answer keys. There were 24 question indicators based on the learning objectives and science process skill indicators. The purpose of making question indicators is to ensure that all items produced are in accordance with the learning objectives and indicators of science process skills. Third, make an assessment rubric. Fourth, conduct a logical validity test to three chemistry lecturers from FMIPA Padang State University and two chemistry teachers. The assessment included 12 criteria that were reviewed from four aspects, namely the suitability of the material, construct, language and additional rules. The logical validity of the instrument was analyzed using Minifac software.

The results of the validator assessment can also be seen in the Measurement Report Expert table. The reliability value of the logical validity test is obtained at 0.82, which is included in the good category according to the Rasch model (Sumintono & Widhiarso, 2015). Furthermore, the exact agrrements score was obtained at 97.2%, which is not much different from the expected agreements value of 97.1%. This means that there is a match between the results of the validator's assessment and the results predicted by the model so that the assessment instrument can be said to be valid (Sick, 2013). Table 1 shows a summary of the results of the logical validity test by the validator.

Table 1. Summary of Logical Validity Test Result

| **Reliability** | **Exact Agreements** | **Expected Agreements** |
|---|---|---|
| 0,82 | 97,2 | 97,1 |

### *Conducting a Pilot Test or Field Test*

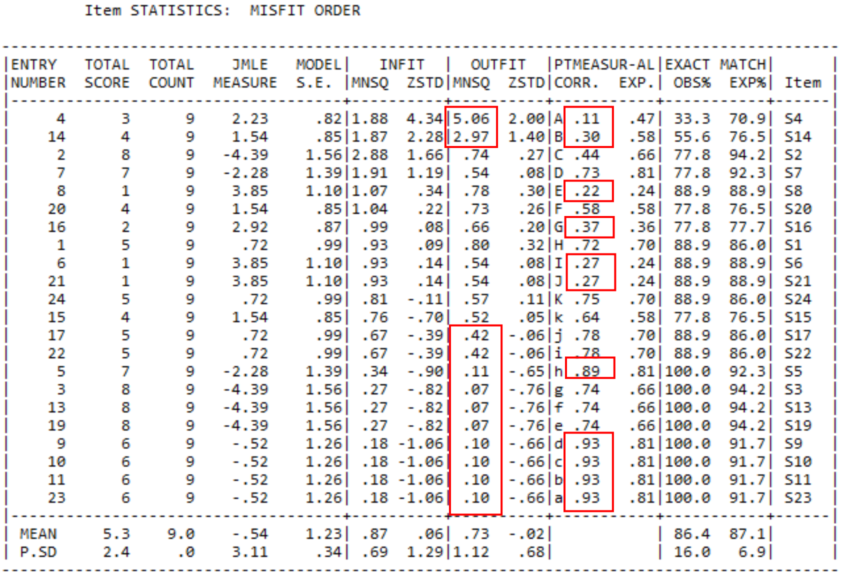
The validated questions should be piloted first with a small number of selected subjects before the field test is conducted. This trial involved nine students in class XI. The sample selection was carried out using purposive sampling method, which means taking samples based on certain considerations (Sugiyono, 2013). In this study, the consideration taken was based on the ability of students, namely high, medium, and low abilities. The goal is to see if all students' abilities can represent the question items. The trial was conducted for 60 minutes, after which an interview was conducted with students which aims to determine the level of understanding of the question items. Before the trial, students were reminded of the stoichiometry material. After the trial, the data obtained was then analyzed using the Rasch model.

### *Conducting Rasch Analysis*

The raw data obtained from the trial results were then analyzed with the Rasch model using Ministep software. The quality of the items reviewed include:

#### a. Validity

Item validity can be analyzed from the item fit menu. According to Boone et al. (2014) and Bond & Fox (2015) the criteria that can be used to analyze item fit are the outfit MNSQ, ZSTD, and Pt Mean Corr values. Of the three criteria, not all of them have to meet the “accepted” value for a question to be said to be valid. If only one criterion is met, the item can still be said to be valid. The question items need to be revised or replaced if the question items on the three criteria are not met (Sumintono & Widhiarso, 2015). Figure 1 shows the results of the validity analysis of the pilot test.



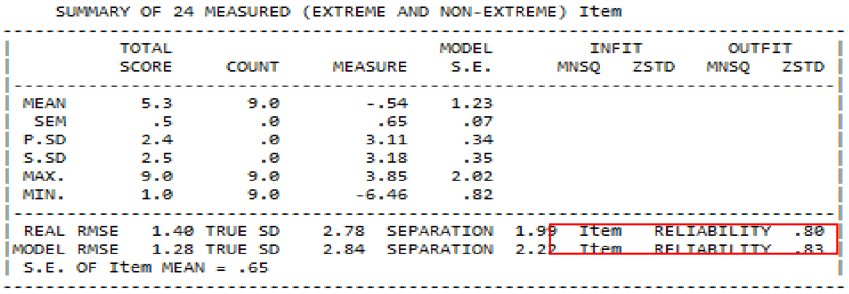
```
        Item STATISTICS:  MISFIT ORDER

------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL   JMLE    MODEL|   INFIT  |  OUTFIT  |PTMEASUR-AL|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item |
|------------------------------------+----------+----------+-----------+-----------+------|
|     4      3      9    2.23     .82|1.88  4.34|5.06  2.00|A .11   .47| 33.3  70.9| S4   |
|    14      4      9    1.54     .85|1.87  2.28|2.97  1.40|B .30   .58| 55.6  76.5| S14  |
|     2      8      9   -4.39    1.56|2.88  1.66| .74   .27|C .44   .66| 77.8  94.2| S2   |
|     7      7      9   -2.28    1.39|1.91  1.19| .54   .08|D .73   .81| 77.8  92.3| S7   |
|     8      1      9    3.85    1.10|1.07   .34| .78   .30|E .22   .24| 88.9  88.9| S8   |
|    20      4      9    1.54     .85|1.04   .22| .73   .26|F .58   .58| 77.8  76.5| S20  |
|    16      2      9    2.92     .87| .99   .08| .66   .20|G .37   .36| 77.8  77.7| S16  |
|     1      5      9     .72     .99| .93   .09| .80   .32|H .72   .70| 88.9  86.0| S1   |
|     6      1      9    3.85    1.10| .93   .14| .54   .08|I .27   .24| 88.9  88.9| S6   |
|    21      1      9    3.85    1.10| .93   .14| .54   .08|J .27   .24| 88.9  88.9| S21  |
|    24      5      9     .72     .99| .81  -.11| .57   .11|K .75   .70| 88.9  86.0| S24  |
|    15      4      9    1.54     .85| .76  -.70| .52   .05|k .64   .58| 77.8  76.5| S15  |
|    17      5      9     .72     .99| .67  -.39| .42  -.06|j .78   .70| 88.9  86.0| S17  |
|    22      5      9     .72     .99| .67  -.39| .42  -.06|i .78   .70| 88.9  86.0| S22  |
|     5      7      9   -2.28    1.39| .34  -.90| .11  -.65|h .89   .81|100.0  92.3| S5   |
|     3      8      9   -4.39    1.56| .27  -.82| .07  -.76|g .74   .66|100.0  94.2| S3   |
|    13      8      9   -4.39    1.56| .27  -.82| .07  -.76|f .74   .66|100.0  94.2| S13  |
|    19      8      9   -4.39    1.56| .27  -.82| .07  -.76|e .74   .66|100.0  94.2| S19  |
|     9      6      9    -.52    1.26| .18 -1.06| .10  -.66|d .93   .81|100.0  91.7| S9   |
|    10      6      9    -.52    1.26| .18 -1.06| .10  -.66|c .93   .81|100.0  91.7| S10  |
|    11      6      9    -.52    1.26| .18 -1.06| .10  -.66|b .93   .81|100.0  91.7| S11  |
|    23      6      9    -.52    1.26| .18 -1.06| .10  -.66|a .93   .81|100.0  91.7| S23  |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN     5.3    9.0    -.54    1.23| .87   .06| .73  -.02|           | 86.4  87.1|      |
| P.SD     2.4     .0    3.11     .34| .69  1.29|1.12   .68|           | 16.0   6.9|      |
------------------------------------------------------------------------------------------
```

Figure 1. Test Validity Analysis Result

Based on Figure 1, of the 24 items developed, only 22 items display the results of their validity analysis on the item fit menu. While the other two items (S12 and S18) do not display the results of their analysis on the item fit menu. Therefore, it can be ascertained that the two items are not good so they need to be revised or replaced. From the results of the analysis of 22 items on the item fit menu, it shows that all items successfully meet the criteria for the outfit ZSTD value with a range of -2.0 to +2.0. This indicates that the item has a logical estimate (Sumintono & Widhiarso, 2015). Furthermore, items S4, S14, S9, S10, S11 and S23 only fulfill one of the three expected criteria, namely the ZSTD outfit value. However, these six items can still be retained. The other minimum criteria met were two of the three fit criteria on each question. The most difficult validity criterion for all questions to achieve was the MNSQ outfit. Although there are 12 questions that are less than the acceptable limit for outfit MNSQ, the outfit ZSTD and Pt Mean Corr values have been met by other items. So based on the Rasch model analysis, it can be concluded that 22 items can be said to be fit. This means that the instrument used measures according to the predetermined objectives (Sumintono & Widhiarso, 2014).

## b. Reliability

Item reliability can be analyzed on the summary statistic menu. Figure 2 shows the results of the reliability analysis of the pilot test.



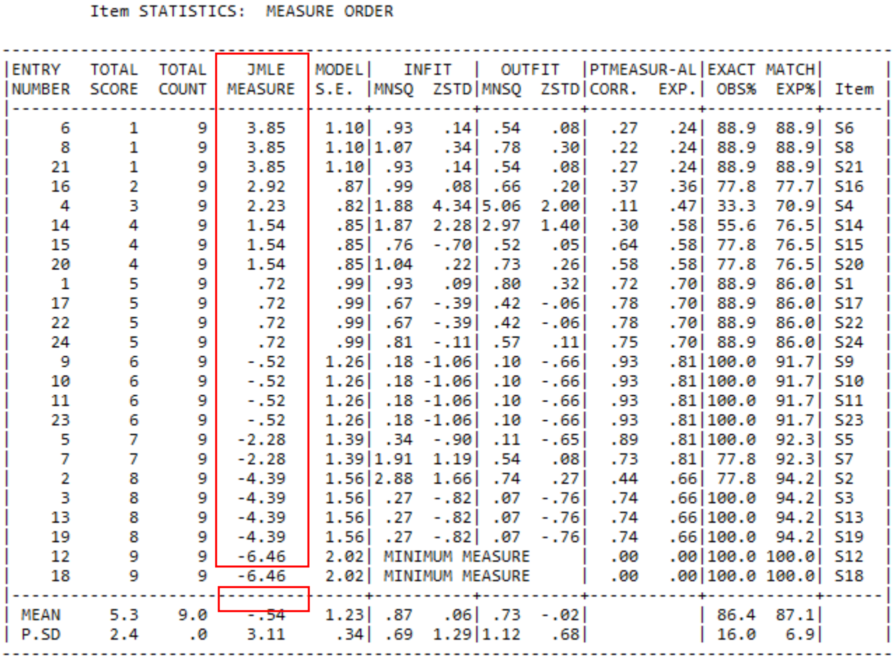
```
      SUMMARY OF 24 MEASURED (EXTREME AND NON-EXTREME) Item
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT     |
|          SCORE      COUNT     MEASURE   S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|------------------------------------------------------------------------------|
| MEAN       5.3        9.0        -.54   1.23                                 |
|  SEM        .5         .0         .65    .07                                 |
| P.SD       2.4         .0        3.11    .34                                 |
| S.SD       2.5         .0        3.18    .35                                 |
| MAX.       9.0        9.0        3.85   2.02                                 |
| MIN.       1.0        9.0       -6.46    .82                                 |
|------------------------------------------------------------------------------|
| REAL RMSE    1.40 TRUE SD     2.78  SEPARATION  1.99  Item    RELIABILITY  .80 |
|MODEL RMSE    1.28 TRUE SD     2.84  SEPARATION  2.22  Item    RELIABILITY  .83 |
| S.E. OF Item MEAN = .65                                                      |
-------------------------------------------------------------------------------
```

Figure 2. Test Reliability Analysis Result

Based on Figure 2, the item reliability value is obtained at 0.80, meaning that the resulting instrument is included in the sufficient category according to the Rasch model. This shows that if the test is repeated over a long period of time, the results obtained for each item will not be much different. Thus, the resulting test instrument is reliable (Sumintono & Widhiarso, 2015).

## c. Difficult Index

The difficulty index can be analyzed on the Item Measure menu. The score to note is located in the JMLE Measure column. The item measure menu also has information about the standard deviation (SD) value. The level of difficulty of the items can be grouped by combining the standard deviation value with the average logit value. Items with a score of 0.00 logit+1SD are categorized as difficult, items with a score of >+1SD are categorized as very difficult, items with a score of 0.00 logit-1SD are categorized as easy, and items with a score of <-1SD are categorized as very easy. Figure 3 shows the results of the pilot test difficulty index analysis.



```
         Item STATISTICS:  MEASURE ORDER

------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL     JMLE   MODEL|   INFIT    |  OUTFIT   |PTMEASUR-AL|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item |
|------------------------------------+----------+----------+-----------+-----------+------|
|     6      1      9     3.85   1.10| .93   .14| .54   .08|  .27   .24| 88.9  88.9| S6   |
|     8      1      9     3.85   1.10|1.07   .34| .78   .30|  .22   .24| 88.9  88.9| S8   |
|    21      1      9     3.85   1.10| .93   .14| .54   .08|  .27   .24| 88.9  88.9| S21  |
|    16      2      9     2.92    .87| .99   .08| .66   .20|  .37   .36| 77.8  77.7| S16  |
|     4      3      9     2.23    .82|1.88  4.34|5.06  2.00|  .11   .47| 33.3  70.9| S4   |
|    14      4      9     1.54    .85|1.87  2.28|2.97  1.40|  .30   .58| 55.6  76.5| S14  |
|    15      4      9     1.54    .85| .76  -.70| .52   .05|  .64   .58| 77.8  76.5| S15  |
|    20      4      9     1.54    .85|1.04   .22| .73   .26|  .58   .58| 77.8  76.5| S20  |
|     1      5      9      .72    .99| .93   .09| .80   .32|  .72   .70| 88.9  86.0| S1   |
|    17      5      9      .72    .99| .67  -.39| .42  -.06|  .78   .70| 88.9  86.0| S17  |
|    22      5      9      .72    .99| .67  -.39| .42  -.06|  .78   .70| 88.9  86.0| S22  |
|    24      5      9      .72    .99| .81  -.11| .57   .11|  .75   .70| 88.9  86.0| S24  |
|     9      6      9     -.52   1.26| .18 -1.06| .10  -.66|  .93   .81|100.0  91.7| S9   |
|    10      6      9     -.52   1.26| .18 -1.06| .10  -.66|  .93   .81|100.0  91.7| S10  |
|    11      6      9     -.52   1.26| .18 -1.06| .10  -.66|  .93   .81|100.0  91.7| S11  |
|    23      6      9     -.52   1.26| .18 -1.06| .10  -.66|  .93   .81|100.0  91.7| S23  |
|     5      7      9    -2.28   1.39| .34  -.90| .11  -.65|  .89   .81|100.0  92.3| S5   |
|     7      7      9    -2.28   1.39|1.91  1.19| .54   .08|  .73   .81| 77.8  92.3| S7   |
|     2      8      9    -4.39   1.56|2.88  1.66| .74   .27|  .44   .66| 77.8  94.2| S2   |
|     3      8      9    -4.39   1.56| .27  -.82| .07  -.76|  .74   .66|100.0  94.2| S3   |
|    13      8      9    -4.39   1.56| .27  -.82| .07  -.76|  .74   .66|100.0  94.2| S13  |
|    19      8      9    -4.39   1.56| .27  -.82| .07  -.76|  .74   .66|100.0  94.2| S19  |
|    12      9      9    -6.46   2.02| MINIMUM MEASURE     |  .00   .00|100.0 100.0| S12  |
|    18      9      9    -6.46   2.02| MINIMUM MEASURE     |  .00   .00|100.0 100.0| S18  |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN     5.3    9.0     -.54   1.23| .87   .06| .73  -.02|           | 86.4  87.1|      |
| P.SD     2.4     .0     3.11    .34| .69  1.29|1.12   .68|           | 16.0   6.9|      |
------------------------------------------------------------------------------------------
```

Figure 3. Test Difficult Index Analysis Results

Based on Figure 3, the SD value is 3.11. Thus, items with logit >3.11 are very difficult questions (S6, S8, and S21). Items with logit 0.0 to 3.11 are difficult questions (S24, S22, S17, S1, S20, S15, S14, S4, and S16). Items with a logit of 0.0 to -3.11 are easy questions (S9, S10, S11, S23, S5, and S7). Items with logit <-3.11 are very easy questions (S2, S3, S13, S19, S12 and S18).

### d. Question Differentiation

The quality of the instrument in terms of overall respondents and items is getting better if the separation value is getting bigger because the instrument can distinguish groups of respondents and items (Sumintono & Widhiarso, 2015). The separation value of the instrument was found to be 1.99 so that the strata value was obtained as follows:

$$H = \frac{(4\,x\,separation) + 1}{3}$$

$$H = \frac{(4\,x\,1{,}99) + 1}{3}$$

$$H = \frac{8{,}96}{3} = 2{,}99$$

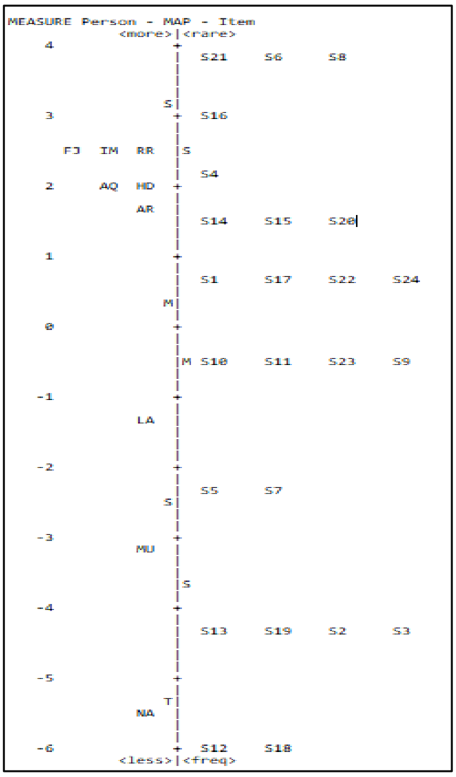
The number 2.99 is rounded to 3, indicating that the instrument can distinguish three groups of items, namely difficult, medium, and easy questions (Sumintono & Widhiarso, 2015).

### *Reviewing Item Suitability Statistics and Revising Items If Necessary*

Based on the results of Rasch analysis, out of 24 items, only 22 items met the validity criteria. Meanwhile, the other two items (S12 and S18) did not meet the expected validity criteria, so these two items needed to be revised first. The reliability value was obtained at 0.80, including in the sufficient category. The questions have a level of difficulty that varies from very difficult questions to very easy questions. Furthermore, the differentiation analysis shows that the instrument can distinguish three groups of items.

### *Viewing the Wright Map*

Wright's map illustrates how learners abilities and item difficulty levels are distributed. The distribution of the level of difficulty of the question is described on the right wright map. Meanwhile, the distribution of students ability levels is described on the left wright map (Sumintono & Widhiarso, 2015). Logit values can be said to be outliers if they are outside the distance of +2SD to -2SD. Figure 4 shows wright map of the pilot test



```
MEASURE Person - MAP - Item
            <more>|<rare>
    4             +
                  |  S21    S6     S8
                  |
                  |
                  |
                 S|
    3             +  S16
                  |
                  |
      FJ  IM  RR  |S
                  |
                  |  S4
    2     AQ  HD  +
                  |
              AR  |
                  |  S14    S15    S20|
                  |
                  |
    1             +
                  |
                  |  S1     S17    S22    S24
                  |
                 M|
                  |
    0             +
                  |
                  |
                  |M S10    S11    S23    S9
                  |
                  |
   -1             +
                  |
              LA  |
                  |
                  |
                  |
   -2             +
                  |
                  |  S5     S7
                 S|
                  |
                  |
   -3             +
              MU  |
                  |
                  |
                  |S
                  |
   -4             +
                  |
                  |  S13    S19    S2     S3
                  |
                  |
                  |
   -5             +
                  |
                 T|
              NA  |
                  |
                  |
   -6             +  S12    S18
            <less>|<freq>
```

Figure 4. Wright Map of Pilot Test

Based on Figure 4, questions S21, S6, and S8 include questions with the highest level of difficulty, namely with a logit of +3.85. This means that the ability of all students to be able to solve this problem correctly is very small. Even so, these three questions are not included in the outlier range (outside the T limits). Meanwhile, items S12 and S18 are the questions with the lowest difficulty level with a logit of -6.46 and are in the outlier range. This is because all students were able to answer both questions correctly so that revisions had to be made first. Furthermore, learners with the code FJ, IM, and RR showed the highest ability with a logit value >+2, where almost answered all item questions correctly and were not in the outlier range. Meanwhile, students with the code NA are students with the lowest ability among other students.

### *If Items Do Not Fit the Rasch Model, Repeat Steps 4-7*

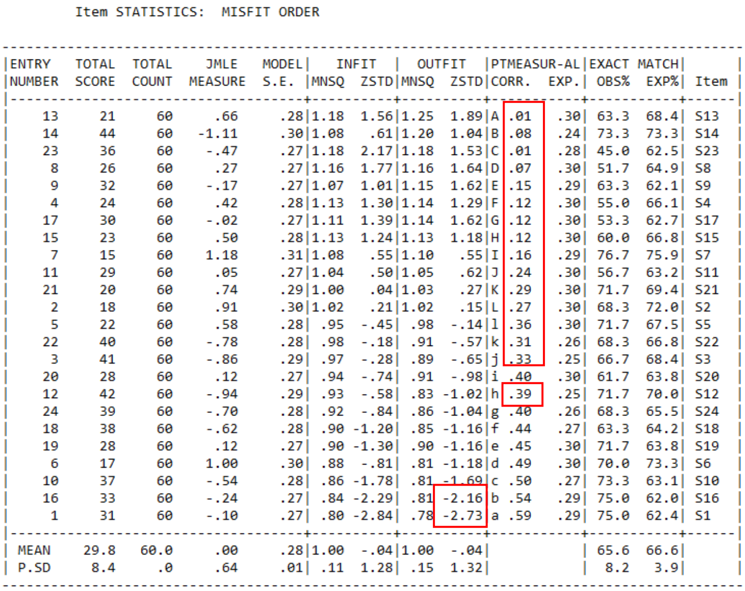
At this stage, steps 4-7 are repeated, because based on the Rasch model analysis there are still 2 items that do not fit and outliers (S12 and S18) so that revisions are needed before the field test is carried out to students. The following steps need to be repeated:

### a. Conduct a Field Test

The field test sample must represent the intended population of 60 students. The sample selection used is adjusted to the number of samples in Rasch modeling, for this number of samples has a confidence level of 99% (Sumintono & Widhiarso, 2014). Before the questions are given, students are reminded of the stoichiometry material. Furthermore, students can work on questions for 60 minutes.

### b. Conducting Rasch Analysis

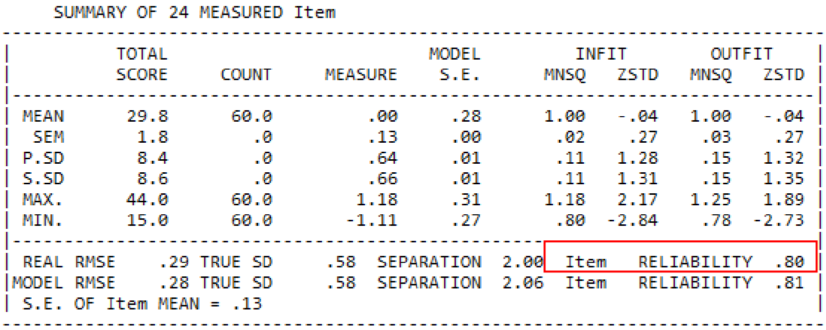
The raw data from the field test were then analyzed with the Rasch model using Ministep software. The quality of the items reviewed included validity, reliability, difficulty index and item differentiation. Figure 5 shows the results of the field test validity analysis.



```
        Item STATISTICS:  MISFIT ORDER

-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL    JMLE    MODEL|   INFIT    |  OUTFIT    |PTMEASUR-AL|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE   S.E. |MNSQ  ZSTD|MNSQ   ZSTD|CORR.   EXP.| OBS%  EXP%| Item |
|-------------------------------------+----------+-----------+-----------+-----------+------|
|    13     21     60      .66     .28|1.18  1.56|1.25  1.89|A .01    .30| 63.3  68.4| S13  |
|    14     44     60    -1.11     .30|1.08   .61|1.20  1.04|B .08    .24| 73.3  73.3| S14  |
|    23     36     60     -.47     .27|1.18  2.17|1.18  1.53|C .01    .28| 45.0  62.5| S23  |
|     8     26     60      .27     .27|1.16  1.77|1.16  1.64|D .07    .30| 51.7  64.9| S8   |
|     9     32     60     -.17     .27|1.07  1.01|1.15  1.62|E .15    .29| 63.3  62.1| S9   |
|     4     24     60      .42     .28|1.13  1.30|1.14  1.29|F .12    .30| 55.0  66.1| S4   |
|    17     30     60     -.02     .27|1.11  1.39|1.14  1.62|G .12    .30| 53.3  62.7| S17  |
|    15     23     60      .50     .28|1.13  1.24|1.13  1.18|H .12    .30| 60.0  66.8| S15  |
|     7     15     60     1.18     .31|1.08   .55|1.10   .55|I .16    .29| 76.7  75.9| S7   |
|    11     29     60      .05     .27|1.04   .50|1.05   .62|J .24    .30| 56.7  63.2| S11  |
|    21     20     60      .74     .29|1.00   .04|1.03   .27|K .29    .30| 71.7  69.4| S21  |
|     2     18     60      .91     .30|1.02   .21|1.02   .15|L .27    .30| 68.3  72.0| S2   |
|     5     22     60      .58     .28| .95  -.45| .98  -.14|l .36    .30| 71.7  67.5| S5   |
|    22     40     60     -.78     .28| .98  -.18| .91  -.57|k .31    .26| 68.3  66.8| S22  |
|     3     41     60     -.86     .29| .97  -.28| .89  -.65|j .33    .25| 66.7  68.4| S3   |
|    20     28     60      .12     .27| .94  -.74| .91  -.98|i .40    .30| 61.7  63.8| S20  |
|    12     42     60     -.94     .29| .93  -.58| .83 -1.02|h .39    .25| 71.7  70.0| S12  |
|    24     39     60     -.70     .28| .92  -.84| .86 -1.04|g .40    .26| 68.3  65.5| S24  |
|    18     38     60     -.62     .28| .90 -1.20| .85 -1.16|f .44    .27| 63.3  64.2| S18  |
|    19     28     60      .12     .27| .90 -1.30| .90 -1.16|e .45    .30| 71.7  63.8| S19  |
|     6     17     60     1.00     .30| .88  -.81| .81 -1.18|d .49    .30| 70.0  73.3| S6   |
|    10     37     60     -.54     .28| .86 -1.78| .81 -1.69|c .50    .27| 73.3  63.1| S10  |
|    16     33     60     -.24     .27| .84 -2.29| .81 -2.16|b .54    .29| 75.0  62.0| S16  |
|     1     31     60     -.10     .27| .80 -2.84| .78 -2.73|a .59    .29| 75.0  62.4| S1   |
|-------------------------------------+----------+-----------+-----------+-----------+------|
| MEAN    29.8   60.0      .00     .28|1.00  -.04|1.00   -.04|           | 65.6  66.6|      |
| P.SD     8.4     .0      .64     .01| .11  1.28| .15   1.32|           |  8.2   3.9|      |
-------------------------------------------------------------------------------------------
```

Figure 5. Field Test Validity Analysis Result

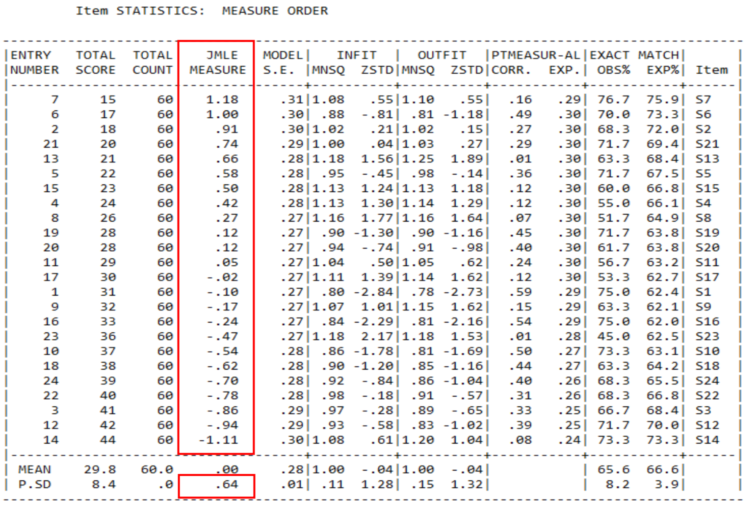
Based on Figure 5, all items successfully meet the MNSQ value criteria with a range of 0.5 - 1.5 which indicates that the measurement on the items is in good condition. The MNSQ value indicates the level of randomness or distortion in the test instrument measurement system. Of the 24 items, there are no items that only meet one criterion. The minimum criteria met by each item were 2 out of 3 criteria. The most difficult validity criterion met by all items is Pt Mean Corr, because of the 24 items there are 16 items that do not meet this criterion. Even so, the MNSQ and ZSTD outfit values have been met by other items. So, based on the Rasch model analysis, all items are suitable (fit) because they have met the expected validity criteria. This means that the instrument used measures in accordance with the objectives that have been set (Sumintono & Widhiarso, 2014). Figure 6 shows the results of the field test reliability analysis.



```
     SUMMARY OF 24 MEASURED Item
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE   S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      29.8       60.0        .00     .28      1.00   -.04   1.00   -.04 |
|  SEM       1.8         .0        .13     .00       .02    .27    .03    .27 |
| P.SD       8.4         .0        .64     .01       .11   1.28    .15   1.32 |
| S.SD       8.6         .0        .66     .01       .11   1.31    .15   1.35 |
| MAX.      44.0       60.0       1.18     .31      1.18   2.17   1.25   1.89 |
| MIN.      15.0       60.0      -1.11     .27       .80  -2.84    .78  -2.73 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .29 TRUE SD     .58  SEPARATION  2.00  Item   RELIABILITY  .80 |
|MODEL RMSE    .28 TRUE SD     .58  SEPARATION  2.06  Item   RELIABILITY  .81 |
| S.E. OF Item MEAN = .13                                                     |
-------------------------------------------------------------------------------
```

Figure 6. Field Test Reliability Analysis Result

Based on Figure 6, the item reliability value is obtained at 0.80, which means that the resulting instrument is included in the sufficient category. This shows that if the test is repeated over a long period of time, the results obtained for each item will not be much different. Thus, the resulting test instrument is reliable (Sumintono & Widhiarso, 2015). Figure 7 shows the results of the field test difficulty index analysis



```
          Item STATISTICS:  MEASURE ORDER

--------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL     JMLE   MODEL|   INFIT    |  OUTFIT   |PTMEASUR-AL|EXACT  MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ   ZSTD|MNSQ   ZSTD|CORR.   EXP.| OBS%   EXP%| Item |
|-------------------------------------+----------+-----------+-----------+------------+------|
|     7     15     60     1.18     .31|1.08    .55|1.10    .55|  .16    .29| 76.7   75.9| S7   |
|     6     17     60     1.00     .30| .88   -.81| .81  -1.18|  .49    .30| 70.0   73.3| S6   |
|     2     18     60      .91     .30|1.02    .21|1.02    .15|  .27    .30| 68.3   72.0| S2   |
|    21     20     60      .74     .29|1.00    .04|1.03    .27|  .29    .30| 71.7   69.4| S21  |
|    13     21     60      .66     .28|1.18   1.56|1.25   1.89|  .01    .30| 63.3   68.4| S13  |
|     5     22     60      .58     .28| .95   -.45| .98   -.14|  .36    .30| 71.7   67.5| S5   |
|    15     23     60      .50     .28|1.13   1.24|1.13   1.18|  .12    .30| 60.0   66.8| S15  |
|     4     24     60      .42     .28|1.13   1.30|1.14   1.29|  .12    .30| 55.0   66.1| S4   |
|     8     26     60      .27     .27|1.16   1.77|1.16   1.64|  .07    .30| 51.7   64.9| S8   |
|    19     28     60      .12     .27| .90  -1.30| .90  -1.16|  .45    .30| 71.7   63.8| S19  |
|    20     28     60      .12     .27| .94   -.74| .91   -.98|  .40    .30| 61.7   63.8| S20  |
|    11     29     60      .05     .27|1.04    .50|1.05    .62|  .24    .30| 56.7   63.2| S11  |
|    17     30     60     -.02     .27|1.11   1.39|1.14   1.62|  .12    .30| 53.3   62.7| S17  |
|     1     31     60     -.10     .27| .80  -2.84| .78  -2.73|  .59    .29| 75.0   62.4| S1   |
|     9     32     60     -.17     .27|1.07   1.01|1.15   1.62|  .15    .29| 63.3   62.1| S9   |
|    16     33     60     -.24     .27| .84  -2.29| .81  -2.16|  .54    .29| 75.0   62.0| S16  |
|    23     36     60     -.47     .27|1.18   2.17|1.18   1.53|  .01    .28| 45.0   62.5| S23  |
|    10     37     60     -.54     .28| .86  -1.78| .81  -1.69|  .50    .27| 73.3   63.1| S10  |
|    18     38     60     -.62     .28| .90  -1.20| .85  -1.16|  .44    .27| 63.3   64.2| S18  |
|    24     39     60     -.70     .28| .92   -.84| .86  -1.04|  .40    .26| 68.3   65.5| S24  |
|    22     40     60     -.78     .28| .98   -.18| .91   -.57|  .31    .26| 68.3   66.8| S22  |
|     3     41     60     -.86     .29| .97   -.28| .89   -.65|  .33    .25| 66.7   68.4| S3   |
|    12     42     60     -.94     .29| .93   -.58| .83  -1.02|  .39    .25| 71.7   70.0| S12  |
|    14     44     60    -1.11     .30|1.08    .61|1.20   1.04|  .08    .24| 73.3   73.3| S14  |
|-------------------------------------+----------+-----------+-----------+------------+------|
| MEAN    29.8   60.0      .00     .28|1.00   -.04|1.00   -.04|           | 65.6   66.6|      |
| P.SD     8.4     .0      .64     .01| .11   1.28| .15   1.32|           |  8.2    3.9|      |
--------------------------------------------------------------------------------------------
```

Figure 7. Result of Field Test Difficulty Index Analysis

Based on Figure 7, the standard deviation value is 0.64. Thus, items with logit >0.64 are very difficult questions (S13, S21, S2, S6, and S7). Items with logit 0.0-0.64 are difficult questions (S11, S20, S19, S8, S4, S15, and S5). Items with a logit of 0.0 to -0.64 are easy questions (S18, S10, S23, S16, S9, S1, and S17). Items with logit <-0.64 are very easy questions (S24, S22, S3, S12 and S14). A good question is one that is not too difficult and not too easy (Asrul et al., 2014). In Rasch modeling there is no medium question category, all question items are categorized into very difficult questions, difficult questions, easy questions, and very easy questions. In essence, the difficulty level of the questions is divided into difficult, medium, and easy questions. Items of difficult questions and easy questions can be said to have a medium difficulty index (Ahmad, 2015). Thus, questions with very difficult categories in the Rasch model include difficult questions, questions with difficult and easy categories include medium questions, and questions with very easy categories include easy questions. Furthermore, the separation value obtained for the analysis of the question's differential power is 2.00 so that the stratum value is obtained as follows:

$$H = \frac{(4\ x\ separation) + 1}{3}$$

$$H = \frac{(4\ x\ 2{,}00) + 1}{3}$$

$$H = \frac{9}{3} = 3$$

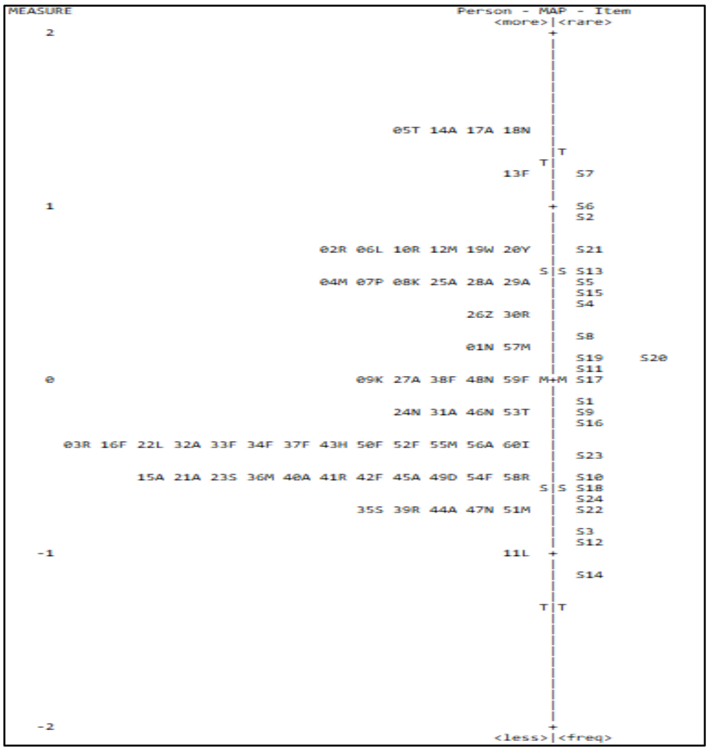
The stratum value is obtained as 3, this indicates that the instrument can distinguish three groups of items, namely difficult, medium, and easy questions (Sumintono & Widhiarso, 2015).

### c. Reviewing Item Suitability Statistics

All items in the field test have met the validity criteria according to the Rasch model. The reliability value obtained is 0.80, including the sufficient category. The analysis of the difficulty index has a variety of question levels ranging from very difficult questions to very easy questions. Furthermore, the analysis of the question's differential power shows that the instrument can distinguish three groups of items

### d. Viewing the Wright Map

Figure 8 shows the Wright map of the field test.



```
MEASURE                                        Person - MAP - Item
                                                    <more>|<rare>
    2                                                     +
                                                          |
                                                          |
                                                          |
                                                          |
                                                          |
                                                          |
                                                          |
                                                          |
                                         05T 14A 17A 18N  |
                                                          |
                                                          |T
                                                         T|
                                                     13F  |  S7
                                                          |
                                                          |
    1                                                     +  S6
                                                          |  S2
                                                          |
                                                          |
                                 02R 06L 10R 12M 19W 20Y  |  S21
                                                          |
                                                         S|S S13
                                 04M 07P 08K 25A 28A 29A  |  S5
                                                          |  S15
                                                          |  S4
                                                 26Z 30R  |
                                                          |
                                                          |  S8
                                                 01N 57M  |
                                                          |  S19      S20
                                                          |  S11
    0                                80K 27A 38F 48N 59F M+M S17
                                                          |
                                                          |  S1
                                         24N 31A 46N 53T  |  S9
                                                          |  S16
                                                          |
     03R 16F 22L 32A 33F 34F 37F 43H 50F 52F 55M 56A 60I  |
                                                          |  S23
                                                          |
             15A 21A 23S 36M 40A 41R 42F 45A 49D 54F 58R  |  S10
                                                         S|S S18
                                                          |  S24
                                     35S 39R 44A 47N 51M  |  S22
                                                          |
                                                          |  S3
                                                          |  S12
   -1                                                11L  +
                                                          |
                                                          |  S14
                                                          |
                                                          |
                                                         T|T
                                                          |
                                                          |
                                                          |
                                                          |
                                                          |
                                                          |
                                                          |
                                                          |
                                                          |
                                                          |
   -2                                                     +
                                                    <less>|<freq>
```

Figure 8. Field Test Wright Map

Based on Figure 8, item S7 is the item with the highest difficulty level with a logit of +1.08 because it occupies the topmost part and is not included in the outlier range. Meanwhile, item S14 occupies the lowest part with the lowest logit value of -1.11 and is not included in the outlier range. Furthermore, students with codes 05T, 14A, 17A and 18N are students with the highest ability. Meanwhile, learners with code 11L are learners with the lowest ability among other learners.

From the Wright map analysis, there are still two questions with the same logit value (S19 and S20). This indicates that the two questions have the same level of difficulty. Even so, the two items do not need to be revised because overall the items are well distributed and there are no outlier items. This can be seen from each interval in the item map being represented by a test item (Sumintono & Widhiarso, 2015).

### *Determining the Instrument Quality Claim*

All items can be claimed valid because they have met the expected validity criteria (outfit MNSQ, ZSTD, and Pt Mean Corr). The reliability of the test instrument is obtained at 0.80, which means that the reliability of the items is included in the sufficient category. The analysis of the difficulty index shows that there are four variations of questions, namely very difficult questions, difficult questions, easy questions, and very easy questions. Then, the question's differential power shows that the instrument can distinguish three groups of items. So, it can be concluded that the test instrument developed to measure the science process skills of students on stoicimetric material has proven to be of high quality because it has been tested for validity, reliability, has a difficulty index and differentiating power according to the Rasch model analysis (Sumintono & Widhiarso, 2015).

### *Developing Documentation for Test Instrument*

After going through the stages of revision, field testing, and analysis of the quality of the instrument, the instrument can be disseminated or used by others. Documentation must clearly state the purpose of using the instrument, instructions for working on the questions, question indicators, question items, assessment rubrics and guidelines for analyzing the quality of the instrument based on the Rasch model.

## 4. Conclusion

Based on the results of the study, it can be concluded that all questions in the test instrument that has been developed to measure students' science process skills in stoichiometry material for grade XI phase F have met the validity criteria as seen based on the outfit values of MNSQ, ZSTD, and Pt. Mean Corr. Reliability on the test instrument is in the sufficient category, has varying levels of difficulty, namely very difficult questions, difficult questions, easy questions, and very easy questions, and has good question discrimination based on the Rasch model analysis. So that this instrument can be used to assist teachers in measuring students' science process skills in stoichiometry material.

## References

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

Amali, N., & Firman, H. (2024). A Framework Design for Developing and Validating Virtual Test to Assess Science Process Skills in Chemistry. *International Conference On Mathematics And Science Education*, 269–279. https://doi.org/10.18502/kss.v9i8.15557

Asmalia, I., Fadiawati, N., & Kadaritna, N. (2015). Pengembangan Instrumen Asesmen Berbasis Keterampilan Proses Sains Pada Materi Stoikiometri. *Jurnal Pendidikan Dan Pembelajaran Kimia*, *4*(1), 299–311.

Asrul, Ananda, R., & Rosnita. (2014). *Evaluasi Pembelajaran.* Medan: Citapustaka Media.

Basuki, F. R., Jufrida, J., Kurniawan, W., Devi, I. P., & Fitaloka, O. (2019). Tes Keterampilan Proses Sains: Multiple Choice Format. *Jurnal Pendidikan Sains (Jps)*, *7*(2), 101–111. https://doi.org/10.26714/jps.7.2.2019.9-19

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Third Edition 3rd Edition*. New York: Routledge.

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

Chetachukwu, B., & Joshua, A. (2022). Assessment of Level of Science Process Skills Possessed by Senior Secondary School Chemistry Students. *AJSTME*, *8*(2), 130–137.

Hamadi, A. A. L., Priyayi, D. F., & Astuti, S. P. (2018). Pemahaman Guru Terhadap Keterampilan Proses Sains (KPS) dan Penerapannya Dalam Pembelajaran IPA SMP di Salatiga. *Jurnal Pendidikan Sains & Matematika*, *6*(2), 42–53. https://doi.org/10.23971/eds.v6i2.935

Kristiyanto, S., Ashadi, Yamtinah, S., & Mulyani, S. (2019). Pengembangan Computerzed Testlet Untuk Mengukur Keterampilan Proses Sains Pada Materi Stoikiometri. *JKPK (Jurnal Kimia Dan Pendidikan Kimia)*, *4*(3), 216–224.

Liu, X. (2020). *Using and Developing Measurement Instruments in Science Education.* New York: Information Age Publishing.

Mahaffy, P. G., Bucat, B., Tasker, R., Kotz, J. C., Treichel, P. M., Weaver, G. C., et al. (2022). *Chemistry Human Activity, Chemical Reactivity.* Toronto: Nelson Education

Mahmudah, L. (2017). Pentingnya Pendekatan Keterampilan Proses Pada Pembelajaran IPA di Madrasah. *ELEMENTARY*, *4*(1), 167–187. https://doi.org/10.21043/elementary.v4i1.2047

Ramadhani, D. K., Susanti, R., & Zen, D. (2015). Pengembangan Soal Keterampilan Proses Sains Pada Pembelajaran Biologi SMA. *Jurnal Pembelajaran Biologi*, *2*, 96–108.

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), 23–26.

Sugiyono. (2013). *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 Assessement 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. https://doi.org/10.1039/c8rp00071a

Trianto. (2009). *Mendesain Model Pembelajaran Inovatif-Progresif.* Jakarta: Kencana.

How to cite this article:

Pertiwi, N., &  Aini, F. Q. (2025). Development of Scientific Process Skill Assessment Instrument on Stoichiometry Using Rasch Model in the Merdeka Curriculum. *Journal of Educational Sciences*, *9(1)*, 241-255.