MANAZHIM
Jurnal Manajemen dan Ilmu Pendidikan

p-ISSN : 2656-5374
e-ISSN : 2656-0216

Terindeks : SINTA 4, DOAJ, Dimensions, Scilit, Lens Crossref, Garuda, Google, etc.

https://doi.org/10.36088/manazhim.v5i2.3800

DEVELOPMENT OF A TEST INSTRUMENT TO TEST STUDENT COMPREHENSION BASED ON MULTIPLE REPRESENTATION OF BUFFER SOLUTION MATERIAL USING THE RASCH MODEL

Kristina Hutagalung & Faizah Qurrata Aini
Universitas Negeri Padang
faizah_qurrata@fmipa.unp.ac.id ; kristinahutagalung07@gmail.com

Abstract

The results of students' daily tests should be an evaluation material for teachers to evaluate students' understanding of the three levels of multiple chemical representations. The use of test instruments can be used as future learning so that teachers can better know which level of representation needs to be emphasized in the learning process and teachers can find out students' understanding abilities of the interconnection of these three levels, especially in buffer solution material. The Rasch model is used in this research, which is of the Research and Development (R&D) variety, to create a test instrument that can assess a student's comprehension of three levels of chemical representation in buffer solution content. The test instrument is valid, reliable, has a difficulty index, and has strong item discrimination. The steps of this research raised 10 stages of development by Wei, et al in 2012. The research instruments used were a questionnaire validating the contents of the items and also a questionnaire validating the assessment rubric which was filled out by 5 chemistry lecturers who are experts in their fields. Raw data from lecturer validation results were analyzed using the Rasch model supported by Minifacet and Ministep Software. The exact agreement results were obtained from the results of the content validity of 90.6% and the expect agreement of 91.1%, which means that there is a fit between the results of the lecturer's assessment and the results of the model design. Furthermore, the results of empirical item quality found that all items met the valid criteria on the MNSQ, ZSTD and Pt-MeanCorr indicators with a reliability value of 0.91 in the very good category, had a difficulty index, namely easy, medium and difficult levels, as well as the differentiating power of the questions. has 3 levels of respondents.

Keywords *: Multiple Representation ; Test Instrument, Buffer Solution, Rasch Model*

INTRODUCTION

The study of matter and how it changes is known as chemistry (Kolomuç & Tekin, 2011). Chemistry is a science that relies on abstract ideas that are challenging for pupils to comprehend, especially when they are asked to believe something they haven't seen (Stojanovska et al., 2017). In chemistry there are complex concepts and phenomena that are abstract and unobservable (Nastiti et al., 2012). Based on the characteristics of chemistry, chemistry will be easy to understand if it is able to be represented into three levels of representation, namely macroscopic, submicroscopic, and symbolic, where these three levels are interrelated and one cannot be ignored (Adadan, 2013). To find out the level of students' understanding of concepts at these three levels, an appropriate assessment system is needed, one of which is the use of a test instrument. Instruments are also useful for assisting educators in knowing how students develop during the learning process (Ropii & Fahrurrozi, 2017).

Based on the results of interviews and questionnaires in three schools, namely Padang 3 SMAN, Padang 10 SMAN, and Padang Laboratory Development High School, out of 3 chemistry teachers at school 2 teachers stated that buffer solution material was material that was difficult for students to understand and understand. That is, not all students were able to understand the concept of buffer solution material, as evidenced by the results of distributing questionnaires given to 30 students, 66.6% of students stated that buffer solution material was material that was difficult to understand because the concepts in the material were too abstract, and hard to imagine. , many formulas, equations and calculations that are difficult to understand. Some sub-materials that are difficult for students to understand are the working principles of buffer solutions, components of buffer solutions and calculating pH values.

In the chemistry learning process that takes place in the classroom, the teacher has conveyed the buffer solution material with explanations at the three levels of multiple chemical representations, but at the sub-microscopic level the teacher has not explained in detail how chemical phenomena are at the sub-microscopic level, as well as the teaching materials used. used by teachers such as learning videos, and PPT is equipped with explanations at the macroscopic, sub-microscopic, and symbolic levels. However, the evaluation instrument provided by the teacher only involves the macroscopic and symbolic levels without involving interconnections at the three levels of chemical representation. This

is not in line, therefore the existing evaluation instruments in schools are not yet effective for use in evaluating students' understanding at these three levels. Even though the results of daily tests should be evaluation material for teachers to evaluate students' understanding of these three levels in a comprehensive and directed manner, so that this test instrument can also be used as future learning so that teachers can better know which level of representation needs to be emphasized in learning, especially in material buffer solution.

Because there is no test instrument to test student understanding at the three levels of representation, a test instrument was developed that can be used to comprehensively test student understanding of buffer solutions. In order to produce good and quality test instruments in terms of validity, reliability, difficulty index, and discriminating power, an analysis of the Rasch model is needed which can provide information about the characteristics of the items in the developed test instrument. The Rasch model was chosen because it has the advantage of more accurately accommodating measurement objectives (Bohori & Liliawati, 2019).

METHODS

Using the Rasch model, this kind of study is known as research and development (R & D). The objectives of this study are to improve upon the 10 stages for creating a test instrument described by Wei et al. in 2012. The 10 steps of development are: (1) establishing the construct, (2) identifying the specified construct, (3) determining the item result space, (4) conducting trials (pilot tests), (5) using the Rasch model to analyze data, (6) evaluating item fit, (7) evaluating the Wright Map, (8) repeating steps 4–7 until it fits the model, (9) establishing claims, and (10) documenting (Wei et al., 2012).

The subjects of this study were 30 grade 11 students from SMAN 3 Padang who had studied buffer solutions and five chemistry teachers from FMIPA UNP. This study's goal is to assess the test instrument's quality in terms of its reliability, validity, difficulty index, and question-creation capability. The type of data used in this study is primary data, which was gathered directly from students and professionals doing content validity checks. The study tool employed was a content validity questionnaire with test items for the buffer solution instrument and a Guttman scale with a "Yes" or "No" response option. With the aid of the minifacet and ministep programs, the Rasch model was used to analyze the study data.

The four criteria—validity, dependability, difficulty index, and item discriminatory power—show that the items meet high standards. The Outfit Mean Square (MNSQ) value, which accepts values between 0.5 and 1.5, the Outfit Z-standard (ZSTD) value, which accepts values between -2.0 and +2.0, and the Point Measure value Correlation (Pt Mean Corr), which accepts values between 0.4 and 0.85, are all used to evaluate the validity of the items. In measuring the reliability of the items, it is based on the value of item reliability with a good category ≥ 8.0. Difficulty index analysis provides information on the standard deviation and logit values for each item which can be seen from the distribution of the difficulty levels of the questions on the Wright Map and in particular can be seen on the Output item measure. The discriminating power of the questions can be analyzed from the separation value. Separation value can be determined by the following equation to see more thoroughly.

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

If the separation value is getting bigger, then the quality of the differentiating power of all items and respondents is also getting better, because it can distinguish between groups of items/items and respondents.

RESULTS

1. *Defining Constructs*

In this first stage, namely making learning progress (Learning Progression) which includes an analysis of learning outcomes and learning objectives from the buffer solution material. From the Learning Outcomes and Learning Objectives that have been analyzed then determine 3 levels of chemical representation, namely macroscopic, sub-microscopic and symbolic representation of each of the Learning Objectives that have been determined. Analysis of learning progression can be seen in Table 1.

Table.1. *Learning Progression*

Learning Objectives (TP)	Cognitive Level	Representasion
Describe the different types of buffer solutions	C2	***Macroscopik :*** when a buffer solution is added acid or base, the pH is relatively constant, or from a pH range, an acidic buffer solution pH 6 (below pH 7) is added a strong acid/strong base the pH becomes 6.5. It hasn't changed much. This can be seen by using indicator paper. ***Sub microscopic:*** there are 2 types of buffer solutions, namely: 1. Supporting acids: weak acids and their conjugate bases, 2. Supporting bases: weak bases and their conjugate acids. ***Symbolic:*** Writing a buffer solution based on its components. For example, the constituent particle acid buffer solution is a weak acid and a conjugate base (acetic acid with acetate ion: CH_3COOH/CH_3COO^-)
Describe the components of a buffer solution	C2	***Sub microscopic***: 1. The components of an acid buffer solution: a weak acid with a conjugat e base (eg: CH_3COOH solution + CH_3COONa solution). The buffer components are CH_3COOH as a weak acid and CH_3COO^- as a conjugate base. 2. Components of a basic buffer solution: a weak base with a conjugate acid (eg: NH_3 solution + NH_4Cl solution). The buffer components are NH_3 as a weak base and NH_4^+ as a conjugate acid.
Explain the working principle of buffer solutions	C2	***Macroscopic :*** When a strong acid/base buffer is added or dilution is added, the pH of the buffer solution is relatively constant. can be seen using indicator paper. For example, an acidic buffer solution has an initial pH of 5, then a strong base is added to 5.1. This proves that the solution can maintain pH with the pH not changing drastically. ***Sub microscopic:*** addition of namesake ions in a weak acid or weak base solution results in a shift in the equilibrium towards the undissociated acid or base molecule. If an acid is added to a buffer solution, the H^+ ions will be bound by the conjugate base in the buffer solution. If what is added is a base, the OH^- ions will be neutralized by the acid in the buffer. - **Acid buffer:** the constituent of acid buffer is a weak acid with a conjugate base. In an acid buffer, it is the conjugate base that plays a role in maintaining the pH.

		- **Buffer base:** the constituent base buffer is a weak base with a conjugate acid. In basic buffers, the conjugate acid plays a role in maintaining the pH. **Symbolically,** the reaction equation the pH of a buffer solution won't change significantly when a strong acid or basic is introduced. For example, in an acid buffer solution, namely CH_3COOH with its salt CH_3COONa when added acid/H^+ will be neutralized by CH_3COO^- which comes from CH_3COONa. The reaction: $H^+_{(aq)} + CH_3COO^- \rightarrow CH_3COOH$
Describe the function of buffer solutions in living organisms and industry.	**C2**	***Sub micro :*** In everyday life there is a function of buffer solutions for example buffer solutions in the bodies of living things (in human blood and cells) and also in the industrial field. **In the body of living things (pH condition in human blood):** There are several natural buffer solutions in the body to maintain blood pH stability. This natural buffer system consists of a carbonate buffer solution and a phosphate buffer solution. Blood has a relatively constant pH, which is around 7.4. This is due to the presence of a H_2CO_3 /HCO_3^- buffer system so that even though the blood is ingested by various substances that are both acidic and alkaline, the effect on changes in pH can be neutralized. If the blood enters a substance that is acidic, the H^+ ions from the acid will react with HCO_3^- ions. Conversely, if the blood enters a substance that is alkaline, the OH^- ions from the base will react with H_2CO_3 (carbonic acid).

2. Identifying Defined Constructs

At this stage, namely compiling/making indicator questions with buffer solution material and selecting the type of test to be used. The indicator questions are compiled based on the results of the analysis of the multiple representation aspects of Learning Objectives. The type of test is an essay test with the aim that students are more free to express their ideas and avoid guessing answers, with a total of 16 grids / question indicators.

3. Define Item Result Space

At this stage designing a test instrument consisting of items and an assessment rubric based on the grid/item indicators that have been analyzed previously. There are 8 discourse

questions with 16 item items which are sub-questions of the discourse questions which have connections to the three levels of chemical representation. In this set of test instruments, the components consist of question discourse, sub-questions, and an assessment rubric. The scoring rubric also contains questions answer keys, achievement levels of students' understanding of chemical multirepresentation, and scoring guidelines. The achievement of students' multi-representational understanding levels can be seen in Table 2.

Table 1. Level of understanding of chemical representations.

Understanding level	Information
Levels 1: connecting symbolic understanding with macroscopic (Identify colors, phenomena that occur and; write the symbols of the reaction equation correctly).	Students are able to answer questions at the level of macroscopic and symbolic understanding.
Level 2: understand sub-microscopic understanding with symbolic (Knowing the shape of the structure of atoms, molecules, and ions; write the symbols of the reaction equation correctly).	Students are able to answer questions at the level of sub-microscopic and symbolic understanding.
Level 3: understand and interpret macroscopic, sub-microscopic and symbolic understanding with macroscopic (Identify colors, phenomena that occur and know the shape of atomic, molecular and ionic structures as well as be able to write symbols of reaction equations correctly; state and explain macroscopic phenomena and processes from sub-microscopic).	- Students are able to answer questions at the level of macroscopic, sub-microscopic, and symbolic understanding. - Students are able to answer questions at the level of macroscopic and sub-microscopic understanding.

(Wang et al., 2017)

After that the test instruments that had been developed were validated by 5 validators of chemistry lecturers who were experts, where content validation was carried out using the Guttman scale with the options "Yes" and "No" against the criteria that had been prepared. The results of content validation were analyzed using the Rasch model with the help of the minifacet program. The following results of content validity analysis by experts can be seen in Figure 1.

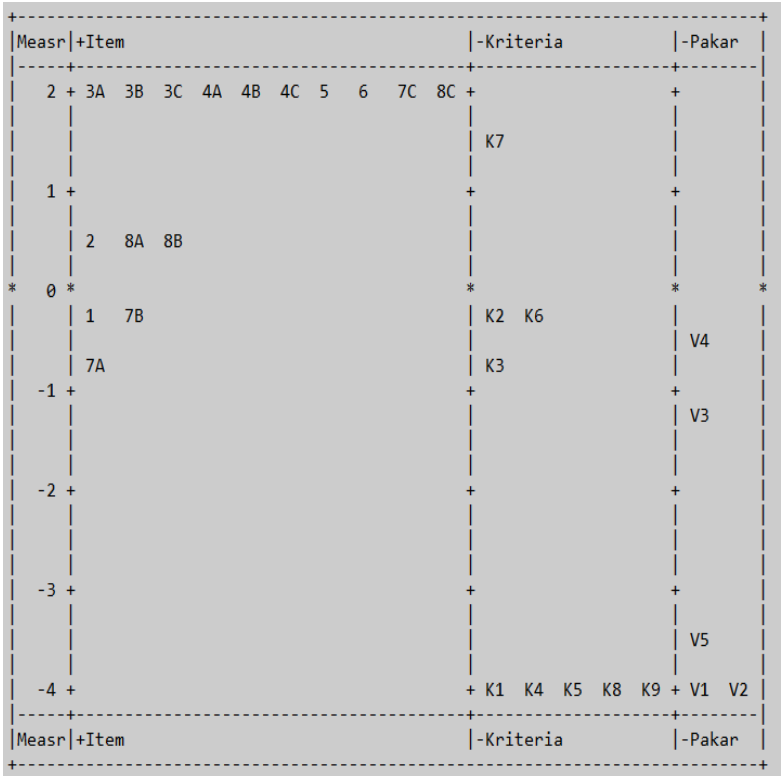

Figure 1. Wright Map

Figure 1 is a Wright map showing the results of content validity data from 5 lecturers' assessment of items with nine assessment aspect criteria. In the picture above there are 4 columns that have different explanations. The first column is the measure column which shows a logit scale with values ranging from -4 to +2. The second column is an explanation regarding the quality of the item from the validator/expert assessment. The way to read it is by looking at the order of the questions from top to bottom, the item that is in the top position means that it has the best quality, because it is able to meet or achieve all the criteria in the assessment aspect. The third column is the criteria for assessing aspects of the test instrument. This section describes the order of the difficulty levels of the aspects. The aspect that is in the top position is the most difficult aspect to be achieved or fulfilled by all items. While the aspects that are at the bottom are the aspects that are easiest to achieve by the item according to the assessment given by the validator. Next, the fourth column, namely the validator column, which explains the order of the validator's assessment.

From the results of content validity using the Rasch model which consisted of 16 essay items, 10 items were obtained with the best quality according to the results of expert assessment, namely question numbers 3A, 3B, 3C, 4A, 4B, 4C, 5, 6, 7C and 8C, where the 10 items of this question are at the very top of the item, which means that they have fulfilled

the assessment aspect, while the item items that are getting lower indicate that the item has not fulfilled the assessment aspect. From the picture above it can also be seen that K7 is the most difficult criterion to fulfill and is in the top position, and the criteria that are getting lower mean that these criteria are easier to achieve.

The summary of the analysis results of the expert assessment of the test instruments analyzed using the Rasch model can be viewed from the four determining indicators as described in Table 2. The first column shows the stratum value obtained, which is 2.90, which indicates that the validator's assessment has met the criteria. reliable. Meanwhile, the validator reliability value obtained was 0.79 with a very good category. This reliability value indicates the reliability of the expert in providing an assessment of the questions (Nisa & Yusmaita, 2022). The third column describes the exact agreement value (validator agreement) with a percentage of 96.2%, which is not much different from the expect agreement (model estimate) with a percentage of 96.1%. This means that there is a fit between the results of the expert assessment and the results predicted by the model (Eliza & Yusmaita, 2021).

Table 2. Summary of expert judgment analysis results (validators)

Strate Value	Reliability	Exact Agreement	Expect Agreement
2,90	0,79	96,2%	96,1%

4. Conducting Trials (test pilots)

Furthermore, the product which has been validated by 5 experts/experts, was tested on 30 class XI students of SMAN 3 PADANG who had studied the buffer solution material. The sample size of 30 met the minimum requirements for testing a description test according to (B. Sumintono & Widhiarso, 2013). Given 16 items of essay sub-questions within 60 minutes for students to answer questions. After the questions have been tested on students, then the answers from students are corrected and given a score against the level of understanding of multiple representations and the scoring guidelines that have been prepared.

5. Analyzing Data with the Rasch Model

The Rasch model was then used to assess the score information gleaned from the students' responses with the aid of the Ministep software. Validity, dependability, difficulty

index, and item discriminatory power were among the qualities of the items under assessment.

1. The validity of the items

The validity of the items in terms of fit order items with 3 fit criteria, namely Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), and Point Measure Correlation (Pt-Mean Corr). According to Boone et al., (2014) in (Planinic et al., 2019) of these three criteria, not all of them must meet the "accepted" value for a question to be said to be "valid". If only one criterion is met, then the question can still be said to be valid. In the results of this analysis there are questions that only meet one criterion. Other minimum criteria that are met are as many as 2 of the 3 fit criteria for each question. Meanwhile, the validity criteria that were the most difficult to achieve were Outfit Pt-Mean Corr. Even though there are 16 questions that are less than the acceptable limit for this Outfit, the Outfit ZSTD and MNSQ Outfit scores can be met by other items. So it can be concluded that all questions can be said to be valid according to the Rasch model analysis with the results that can be seen in Figure 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item	G
5	59	30	-2.34	1.02	1.06	.38	2.01	1.06	A -.10	.10	96.6	96.6	3C	B
15	56	30	-.82	.54	1.23	.66	2.00	1.72	B -.24	.20	86.2	86.4	8B	B
8	55	30	-.55	.50	1.12	.46	1.31	.81	C .01	.22	82.8	83.0	4C	B
11	57	30	-1.15	.62	1.00	.16	1.22	.54	D .13	.17	89.7	89.8	7A	B
13	52	30	.07	.42	1.22	.88	1.13	.50	E .30	.27	79.3	73.1	7C	B
12	58	30	-1.60	.74	.99	.20	1.21	.51	F .11	.14	93.1	93.2	7B	B
1	54	30	-.32	.46	1.12	.47	1.17	.56	G .06	.23	79.3	79.7	1	B
2	54	30	-.32	.46	1.12	.47	1.17	.56	H .06	.23	79.3	79.7	2	B
10	194	30	1.65	.17	.88	-.01	1.13	.40	h .32	.34	65.5	61.8	6	G
6	58	30	-1.60	.74	.98	.18	.80	-.02	g .19	.14	93.1	93.2	4A	B
9	191	30	1.73	.15	.62	-.66	.98	.18	f .37	.36	55.2	59.7	5	G
14	57	30	-1.15	.62	.95	.06	.69	-.38	e .27	.17	89.7	89.8	8A	B
16	37	30	2.03	.34	.88	-.43	.88	-.43	d .43	.39	69.0	63.1	8C	B
3	56	30	-.82	.54	.86	-.22	.62	-.71	c .38	.20	86.2	86.4	3A	B
7	65	30	2.62	.16	.72	-1.27	.82	-.71	b .69	.59	27.6	28.7	4B	D
4	66	30	2.59	.16	.71	-1.31	.80	-.79	a .68	.58	31.0	28.8	3B	D
MEAN	73.1	30.0	.00	.48	.97	.0	1.12	.2			75.2	74.5		
P.SD	45.6	.0	1.55	.24	.17	.6	.39	.7			20.3	20.5		

Figure 2. *Item Fit Order*

2. Item Reliability

Reliability analysis on the *Output Summary Statistics* in the Rasch model provides information on the *Cronbach's Alpha* value of 0.88. *Cronbach's Alpha* value is used to measure overall reliability by looking at the interaction between person and item (Sumintono & Widhiarso, 2015). Reliability explains whether an instrument provides the same or consistent information if repeated measurement tests are carried out (Pratama, 2020). With a Cronbach

Alpha value of > 0.88, it means that the reliability of the test instrument is in the Good category. In addition, to see the reliability of specific items, you can review the item reliability shown in Table 4 with a value of 0.88 which means it has good reliability. The conclusion is that the test instrument can be said to be reliable.

```
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .89
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .45  SEM = 2.99

     SUMMARY OF 16 MEASURED (NON-EXTREME) Item
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      73.1      30.0         .00     .48       .97    .00   1.12    .24 |
|  SEM      11.8        .0         .40     .06       .04    .16    .10    .18 |
| P.SD      45.6        .0        1.55     .24       .17    .62    .39    .68 |
| S.SD      47.0        .0        1.60     .25       .18    .64    .40    .70 |
| MAX.     194.0      30.0        2.62    1.02      1.23    .88   2.01   1.72 |
| MIN.      37.0      30.0       -2.34     .15       .62  -1.31    .62   -.79 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .55 TRUE SD    1.45  SEPARATION  2.65  Item   RELIABILITY  .88 |
|MODEL RMSE    .53 TRUE SD    1.46  SEPARATION  2.73  Item   RELIABILITY  .88 |
| S.E. OF Item MEAN = .40                                                     |
```

Figure 3. *Summary Statistic*

3. Different Power Questions

The differential power of the item or in Rasch's modeling is called item discrimination power, explaining how good the level of the item is to be able to compare individuals who have high (high) and low (low) abilities. The different power of the questions is also analyzed using the Output Summary Statistics in Table 4 above. The grouping of different power items can be viewed from the value of separation. If the separation value is greater, then the quality of the distinguishing power of all items and respondents is also getting better, because it can distinguish groups of questions and respondents (Sumintono & Widhiarso, 2015). With a separation value of 2.65, H = [(4×2.65) + 1])/3 = 3.86. The number 3.86 is rounded to 4. This means that there are 4 groups of item items, namely questions that are easy, medium, difficult, and very difficult.

4. Difficulty Indeks

Analysis of the difficulty index can be viewed from the Output Item Measure which is presented in figure 5. The item difficulty index is obtained from a combination of the mean value and the Standard Deviation (SD) value. The average logit value is 0.00 and the standard deviation is 1.16. Then the average value of the logit measure is 1.16 logit. Then the item difficulty level group is obtained as shown in Figure 2.

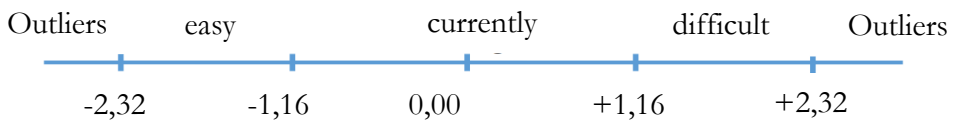

Figure 4. Variation of Item Difficulty Levels

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item	G
7	65	30	2.62	.16	.72	-1.27	.82	-.71	.69	.59	27.6	28.7	4B	D
4	66	30	2.59	.16	.71	-1.31	.80	-.79	.68	.58	31.0	28.8	3B	D
16	37	30	2.03	.34	.88	-.43	.88	-.43	.43	.39	69.0	63.1	8C	B
9	191	30	1.73	.15	.62	-.66	.98	.18	.37	.36	55.2	59.7	5	G
10	194	30	1.65	.17	.88	-.01	1.13	.40	.32	.34	65.5	61.8	6	G
13	52	30	.07	.42	1.22	.88	1.13	.50	.30	.27	79.3	73.1	7C	B
1	54	30	-.32	.46	1.12	.47	1.17	.56	.06	.23	79.3	79.7	1	B
2	54	30	-.32	.46	1.12	.47	1.17	.56	.06	.23	79.3	79.7	2	B
8	55	30	-.55	.50	1.12	.46	1.31	.81	.01	.22	82.8	83.0	4C	B
3	56	30	-.82	.54	.86	-.22	.62	-.71	.38	.20	86.2	86.4	3A	B
15	56	30	-.82	.54	1.23	.66	2.00	1.72	-.24	.20	86.2	86.4	8B	B
11	57	30	-1.15	.62	1.00	.16	1.22	.54	.13	.17	89.7	89.8	7A	B
14	57	30	-1.15	.62	.95	.06	.69	-.38	.27	.17	89.7	89.8	8A	B
6	58	30	-1.60	.74	.98	.18	.80	-.02	.19	.14	93.1	93.2	4A	B
12	58	30	-1.60	.74	.99	.20	1.21	.51	.11	.14	93.1	93.2	7B	B
5	59	30	-2.34	1.02	1.06	.38	2.01	1.06	-.10	.10	96.6	96.6	3C	B
MEAN	73.1	30.0	.00	.48	.97	.0	1.12	.2			75.2	74.5		
P.SD	45.6	.0	1.55	.24	.17	.6	.39	.7			20.3	20.5		

Figure 5. *Item Measure*

Based on the results of the item measure analysis in figure 5, two outlier items were obtained, namely items no. 4B and 3B with a logit value of +2.62 and +2.59 because they exceeded the logit value of +2.32. Besides that, no other outlier items were found, so the difficulty index of the questions can be grouped, consisting of variations that are easy, moderate, difficult, to very difficult (Palimbong et al, 2018).

6. Review Item Compatibility

At this stage, a review of the items is carried out based on the results of the analysis of the four criteria in the previous stage. From the validity analysis, all items are fit and in accordance with the model. From the reliability analysis with an item reliability value of 0.88, it means that the item reliability is good. Difficulty index analysis provides information that there are variations of questions that are easy, medium, difficult, and 2 items are very difficult. For different power the questions have three different power which are considered appropriate and good according to the model.

7. Viewing the Wright Map

The distribution of student/respondent skills and the degree of difficulty of items on the same scale are thoroughly explained by the Wright Map analysis. The distribution of student skills is shown on the left side of the Wright map, while the distribution of item difficulty levels is shown on the right side. There are 2 students with the highest position which means they have the highest ability with logit values > +2, namely 09P and 20P. While students with the lowest ability are in the lowest position, namely 01L and 12P with a logit value of >+1. Besides that, the most difficult item items are items no. 3B and 4B because they occupy the top position and are outside the T limit (outlier). The items with the lowest logit value (≤-2 logit) are item no 3c, but are not included in the outlier range. In this case, it means that students have more correct answers to these questions (Sabekti & Khoirunnisa, 2018). The distribution of students' Wright maps can be seen in Figure 6.

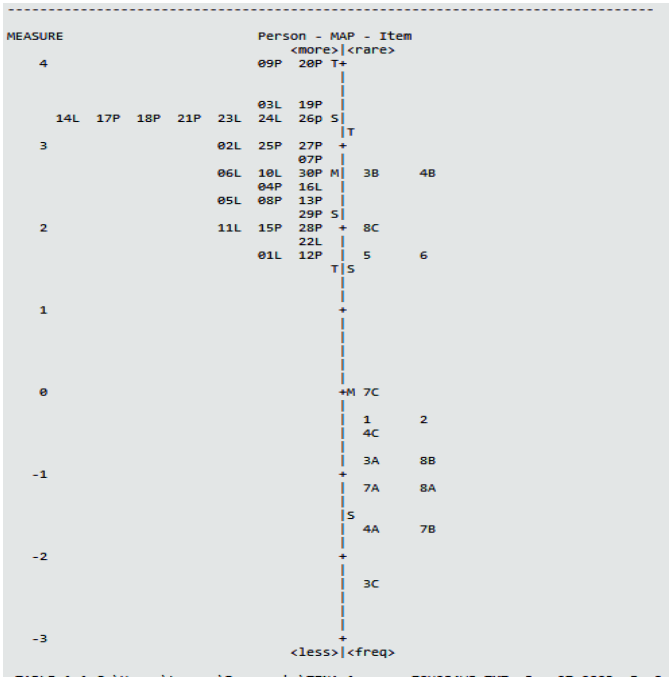

Figure 6. Peta Wright

8. Repeating Steps 4-7 Until All Items Fit

The expected quality of the test instrument has been achieved with evidence of the results of the analysis of validity, reliability, difficulty index and different power of questions that are good and in accordance with the model. However, when viewed from the Wright Map analysis, there are two items that are outside the T (outlier) limits, namely items no. 3B

and 3B. however, numbers 3B and 4B are considered still feasible to use because they are not too far from the previous item and person, namely at logit +2.62 and +2.59. Although this item needs to be considered for revision and re-testing, the second trial and re-testing were not carried out due to limited research time.

9. Establish Test Instrument Quality Claims

The validity of all item items has been detailed in Table 4. All items can be claimed to be valid because they meet at least 2 of the 3 criteria for MNSQ, ZSTD, and Pt-Mean Corr in the Item Fit Order table. As for the reliability aspect of the test instrument, it can be claimed to be reliable based on the Summary Statistics data in Table 5 with the item reliability value = 0.88. Likewise, the difficulty index and different power of the questions obtained have 34 variations of the level of questions from easy, medium to difficult questions.

10. Develop Test Instrument Documentation

Test instrument documentation needs to be developed to provide information to teachers and students in using the test instrument. So that the information obtained is more complex related to the characteristics of the test instrument being developed (Sabekti & Khoirunnisa, 2018). The documents required for this test instrument are learning progression, test instrument grids, items (covering covers and general instructions), assessment rubrics, and guidelines for achieving students' understanding levels in macroscopic, sub-microscopic, and symbolic representations.

DISCUSSION

The resulting test instrument can test the level of understanding of students on acid base materials comprehensively. A student can be said to have understanding level three if it succeeds in achieving a maximum (perfect) score on all sub-items representing each level of representation (Wang et al., 2017). For example, if students succeed in answering the questions correctly and completely which has three levels (macroscopic, sub-microscopic, and symbolic aspects), meaning that the student's understanding is at level 3. However, if the student answered the submicroscopic-symbolic questions correctly, meaning that the understanding was there at level 2. Meanwhile, if students are able to answer the questions correctly macroscopic-symbolic or on one of the questions (macroscopic/submicroscopic/ symbolic), meaning that his understanding is at level 1.

Based on the review of the validity of the items in Figure 2 which displays item fit order, two questions were found that did not meet the criteria for the outfit mean square (0.5 < MNSQ < 1.5), namely items no. 3c and 8b with MNSQ values of 2.01 and 2.00. Statistically, the mean square value is the statistical value of chi square divided by the degrees of freedom, so the value is always positive (Sumintono & Widhiarso, 2013). The MNSQ value serves to indicate the suitability of the data with the model. Here, means that item no. 3c and 8b are considered less productive for measurement, may cause errors with high reliability values. However, this does not degrade the quality of the test instrument because of the criteria ZSTD is fit, even though the Pt-Meean Corr is not fit, but not all of them must meet the "accepted" score for a question to be said to be "valid". If only one criterion is met, then the question can still be said to be valid. Likewise with the Pt-Mean Corr score, there are 14 items that are not fulfill (<0.4), namely item no 3c, 8b, 4c, 7a, 7c, 7b, 1, 2, 6, 4a, 5, 8a, 8c, and 3a. This means that the items tend not to fit the score Acquired Pt-Mean Corr. However, the item can be maintained because of two other criteria have been fit. Pt-Mean Corr provides grain presence information misleading questions (when subjects with low ability answered correctly, while high ability answered incorrectly).

Meanwhile, ZSTD serves as a t-test showing the hypothesis suitability of the data with the model (Sumintono & Widhiarso, 2015). based on data obtained, it was found that all items met the ZSTD criteria. This means that there are no item deviations and all items are considered appropriate with models. The validity of the instrument as a whole can be seen from the average value average MNSQ and ZSTD outfits. This test instrument has an average (mean) of 1.12 and an average ZSTD of 0.2. Based on these results, it means the test instrument declared valid and can measure aspects that are in accordance with the objectives measurement as it should (Sumintono & Widhiarso, 2015)

The resulting test instrument has fulfilled the valid criteria, reliability, has a difficulty index and has good item discrimination. Even though this test instrument has been claimed to be valid, reliable, has an index difficulty and different power of good questions, but based on the analysis, in fact these results cannot be used as a reference to determine students' level of understanding at the interconnection of macroscopic, sub-microscopic, and symbolic. The minimum score obtained should be increasing from level 1 to level 3, but in fact the minimum score is at level 1 (macroscopic-symbolic interconnection, and symbolic-symbolic) is higher than level 2 and 3. Therefore it needs to be reconsidered to do re-testing (pilot test) of the subject, in order to obtain consistent results It is hoped that a guide can be developed

in determining the level understanding of students in general and can diagnose the location of student weaknesses.

CONCLUSION

Based on the study's findings, it can be said that the test materials created to evaluate students' comprehension of the macroscopic, sub-microscopic, and symbolic levels in acid-base materials have satisfied the requirements of being valid, reliable, having a good index of difficulty, and having the ability to differentiate between students. The value of the exact agreement (96.2%) and the expect agreement (96.1%) have not too much difference, meaning that there is a fit between the results of the approval of the validity by the experts and the model estimates. The results of the validity test for students have met the criteria of MNSQ, ZSTD, and Pt-Mean Corr. The item reliability of 0.88 means good. The index of difficulty and discriminating power of the items analyzed has three variations of items from easy, medium, and difficult. The quality test results of the test instruments empirically meet the criteria of MNSQ, ZSTD, and Pt-Mean Corr on the element of validity. The item reliability of 0.88 means good. The difficulty index has a variety of question levels from easy (12.5%), medium (75%), difficult (8.3%) and very difficult (4.2%). Finally, the discriminating power of the questions was recorded with a separation value (H) = 4, which means that the test instrument is able to distinguish students with low, medium and high abilities.

REFERENCES

Adadan, E., 2013, 'Using multiple representations to promote grade 11 students' scientific understanding of the particle theory of matter', Research Science Education, Vol. 43, hh. 1079-1105.

Bohori, M., & Liliawati, W. (2019). Analisis penguasaan konsep siswa menggunakan Rasch Model pada materi usaha dan energi. *Prosiding Seminar Nasional Fisika*, *0*, 138–143. http://proceedings.upi.edu/index.php/sinafi/article/view/579

Eliza, W., & Yusmaita, E. (2021). Pengembangan Butir Soal Literasi Kimia pada Materi Sistem Koloid Kelas XI IPA SMA/MA. *Jurnal Eksakta Pendidikan (Jep)*, *5*(2), 197–204. https://doi.org/10.24036/jep/vol5-iss2/621

Kolomuç, A., & Tekin, S. (2011). Chemistry Teachers' Misconceptions Concerning Concept of Chemical Reaction Rate. *International Journal of Physics & Chemistry Education*, *3*(2), 84–101. https://doi.org/10.51724/ijpce.v3i2.194

Nastiti, R. D., Fadiawati, N., Kadaritna, N., & Diawati, C. (2012). *DEVELOPMENT MODULE OF REACTION RATE BASED ON MULTIPLE*

REPRESENTATIONS Ruli Dwi Nastiti 1 , Noor Fadiawati 2 , Nina Kadaritna 2 , Chansyanah Diawati 4 Pendidikan Kimia Universitas Lampung. 1–15.

Nisa, D. Q., & Yusmaita, E. (2022). Pengembangan Butir Soal Literasi Kimia pada Topik Larutan Elektrolit dan Non Elektrolit Kelas X SMA / MA Development of Chemical Literacy Items in Electrolyte and Non Electrolyte Solution Topic for Class X SMA / MA. *Entalpi Pendidikan KImia*, *17*, 49–57.

Physics Education Journal. (2018). Universitas Papua. *Physics Education Journal*, *1*(1), 12–21. i.yusuf@unipa.ac.id

Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research*, *15*(2), 20111. https://doi.org/10.1103/PhysRevPhysEducRes.15.020111

Pratama, D. (2020). Analisis Kualitas Tes Buatan Guru Melalui Pendekatan Item Response Theory (IRT) Model Rasch. *Tarbawy : Jurnal Pendidikan Islam*, *7*(1), 61–70. https://doi.org/10.32923/tarbawy.v7i1.1187

Ropii, M., & Fahrurrozi, M. (2017). Evaluasi Hasil Belajar. Evaluasi Hasil Belajar. In *Yogyakarta: Pustaka Pelajar*.

Sabekti, A. W., & Khoirunnisa, F. (2018). Penggunaan Rasch Model Untuk Mengembangkan Instrumen Pengukuran Kemampuan Berikir Kritis Siswa Pada Topik Ikatan Kimia. *Jurnal Zarah*, *6*(2), 68–75. https://doi.org/10.31629/zarah.v6i2.724

Stojanovska, M., M. Petruševski, V., & Šoptrajanov, B. (2017). Study of the Use of the Three Levels of Thinking and Representation. *Contributions, Section of Natural, Mathematical and Biotechnical Sciences*, *35*(1), 37–46. https://doi.org/10.20903/csnmbs.masa.2014.35.1.52

Sumintono, B. . W. W. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. Trim Komunikata.

Sumintono, B., & Widhiarso, W. (2013). *Aplikasi Model Rasch Untuk Penelitian Ilmu-Ilmu Sosial*.

Wang, Z., Chi, S., Luo, M., Yang, Y., & Huang, M. (2017). Development of an instrument to evaluate high school students' chemical symbol representation abilities. *Chemistry Education Research and Practice*, *18*(4), 875–892. https://doi.org/10.1039/c7rp00079k

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