J. Pijar MIPA, Vol. 18 No. 2, March 2023: 208-218
DOI: 10.29303/jpm.v18i2.4720

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

DEVELOPMENT OF AN ASSESSMENT INSTRUMENT TO EVALUATE THE CONCEPTS OF CHEMICAL EQUILIBRIUM AT MACROSCOPIC, SUB-MICROSCOPIC, AND SYMBOLIC REPRESENTATION LEVELS USING THE RASCH MODEL

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

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

Abstract: This study aims to develop a test instrument that can be used to evaluate the concept of chemical equilibrium at the macroscopic, submicroscopic, and symbolic levels of high school students who are declared valid, reliable, have a difficulty index, and have good item discrimination. This study was designed using the Rasch development model, which Wei has modified with ten research stages. The results of the research to test content validity were analyzed through the minifacets program with five validators consisting of four chemistry lecturers at Padang State University and one teacher at high school SMAN 3 Padang, obtaining fit analysis results seen from the exact agreement value of 89.1%, which is very close to the expected agreement value of 89.7%. The test results on 30 students of SMAN 3 Padang who were analyzed through the ministep program showed valid results for all items because they met the criteria for the MNSQ, ZSTD, and Pt Mean Corr outfit scores. The reliability value obtained for the item is 0.74, which meets the sufficient criteria, and Cronbach's alpha reliability value is 0.74, which meets the good criteria. All question items also do not have DIF based on gender. The analyzed index of difficulty and discriminating power fulfilled the good criteria with the classification of easy, medium, and difficult items.

Keywords: *Test Instruments, Chemical Equilibrium, Multiple Representation, Rasch Models*

INTRODUCTION

Chemistry is a branch of science that is based on concepts and has abstract material, making it difficult for students to learn and understand it, especially if students have to understand something without being able to see it for real.[1]. Chemistry has an abstract concept; the learning process can be supported by applying the three levels of representation[2]. There are three levels of representation, namely the macroscopic level involving real phenomena that students can directly observe through the senses; the symbolic level involves the representation of atoms, ions, and molecules in the form of symbols, graphics, etc.; The submicroscopic or particulate level has the most important role, including in explaining abstract processes at the particulate level, such as atoms, ions, and molecules.

Learning that can connect up to the particulate level (submicroscopic level) with other levels (macroscopic and symbolic levels) will be very effective in helping students make connections between the three levels in understanding chemical material so that the understanding gained by students will be intact and comprehensive.[3]. Based on the results of a preliminary study conducted on 90 students in three schools in Padang, chemical equilibrium is difficult for students to understand. In his learning process, the teacher connected the three levels of representation in chemical equilibrium material. However, the test instrument the teacher uses in evaluating is less effective and not comprehensive in measuring student abilities because it has yet to connect the three levels of representation. The test instruments used by most of the teachers use questions contained in the LKS or printed books which are then modified. The problems used have drawbacks; namely, they are only focused on connecting between the macroscopic and symbolic levels, so the instrument has yet to connect up to the submicroscopic level. Connecting the three levels of representation in chemical equilibrium material will grow students' ability to interpret the submicroscopic level with other levels and explore students' imaginations. Students will be able to use their mental models to explain events through visual models. Their difficulties will help them interpret sub-micro-molecule structures and make students' creativity grow and develop in understanding chemistry comprehensively if they continue to be trained and accustomed to questions that connect the three representative levels.[4].

Based on the problems described, a test instrument was developed that can connect the three levels of representation in chemical equilibrium material. The test instrument developed in this study is in the form of essay questions. Answering essay questions will activate students' higher-order thinking skills in finding solutions and answers rather than just using memory[5]. The ability to solve problems through essay questions also has advantages compared to using multiple-choice questions. The ability to solve problems using essay questions will enable the teacher to see the quality of workmanship and

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

problem-solving answered by students, compared to answers to multiple choice questions, which only see students' final results [6].

The developed test instrument is good if it can reveal data by reality, is permanent, and is proven to have validity, reliability, level of difficulty, and discriminatory power to reveal the quality of the items that have been developed [7]. The Rasch model is an analytical model used to perform instrument analysis using a rating scale measurement. Using the Rasch model, response theory will have latent logistic properties that focus on the quality of outcome measures. The Rasch analysis aims to provide more precise and accurate measurements of people and item items that can support various aspects of validity and precision[8]. The advantages of using the Rasch model are being able to identify error responses, missing data scores, the ability to depend on the results of the correct answer, and identifying if there is a guessing response in the answer[9].

RESEARCH METHODS

This research was conducted at SMAN 3 Padang in the odd semester of the 2022/2023 academic year. The subjects in this study consisted of four chemistry lecturers, one chemistry teacher at SMAN 3 Padang, and 30 grade 12 students at SMAN 3 Padang who had studied chemical equilibrium material.

This type of research is Research and Development (R&D) using the Rasch model research design, which has been modified in its development stage based on the results of Wei's research with ten stages of development, namely, (1) defining the construct, (2) identifying the construct in the form of an indicator question, (3) determine the item result space by compiling item items, scoring rubrics, scoring guidelines, and guidelines for analyzing students' multi-representation understanding levels, (4) conducting instrument trials, (5) analyzing raw data by applying the Rasch model using Ministep software, (6) reviewing the suitability of item items based on the results of the Rasch model analysis and revision of item items if needed, (7) reviewing the Wright Map, (8) repeating steps 4-7 until all item items match the Rasch model,(9) establish claims for validity, reliability, difficulty index, and differential power of each item, and (10) document the instrument[10].

The product developed is in the form of 6 essay questions. The type of data in this research uses primary data because the data is directly obtained from experts/experts and students. The object of this study is the quality of the instrument in terms of validity, reliability, difficulty index, and item discriminating power. The data analysis technique in this study was carried out using the Rasch model with the help of minifacets and ministep software. Data analysis using minifacets was carried out on data obtained during content validation with lecturers and teachers. In the minifacets program, you will see the strata value, exact agreement, and expected agreement. Data analysis processed using the ministep program was carried out on data obtained from students and will be seen from the validity value in terms of MNSQ, ZSTD, and PT Mean Corr criteria. The reliability value can be seen from the value of the item reliability and Alpha Cronbach's. The difficulty and differential power index can be seen from the values in the statistical summary output table.

RESULTS AND DISCUSSION

Defining Constructs

In the first stage, the construct is defined based on understanding the material and linearity. It is done through theoretical considerations with unidimensional properties from a lower level to a higher level research subject. The basic competencies KD that will be used are KD 3.8, explaining the equilibrium reaction in the relationship between the reactants and the products of the reactants at the equilibrium constant, and KD 3.9, analyzing the factors that influence the direction of the equilibrium and its application in industry. The results of the analysis are made in the form of a learning progress table in the form of competency achievement indicators (GPA), cognitive level, and level of representation as follows (table 1):

Identifying Constructs

At this stage, the level of constructs that have been determined after analyzing learning progress will be identified by compiling indicator questions. The question indicators include the subject matter, the form of the questions, and the items.

Perform Item Result Room Determination

At this stage, the items that have been designed are specified, and then the item results are determined. The resulting item space consists of three parts to obtain questions that are feasible to be tested, namely:

1) Designing ratings and assessment rubrics

The analysis results in the item design obtained 6 question indicators and 16 sub-grains of questions. Each item has aspects of macroscopic, sub-microscopic, and symbolic levels that are essential points in making this problem. For the design of the items that have been completed, it is necessary to make an assessment rubric so that the specifications in each item about the subjectivization of the assessment. In addition, so that the determination of representation of each item is clear, a level of understanding of representation is made to see the relationship

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

between the level of representation in the question item. The levels of understanding of multiple chemical representations will be an indicator of achievement for students; categorizing the level of understanding of students can be classified according to the category in Table 2 [11].

Table 1. Learning Progression

Key GPA	Levels cognitive	Representation
3.8.7 Describe the reaction equilibrium in terms of the relationship between the reactants and the products of the reactants at the equilibrium constant	C2	*symbolic*, Equilibrium Reaction Equations
3.9.1 Analyzing the concentration factors that affect the direction of the shift in the equilibrium	C4	*macroscopic*, the color changes when there is an increase/decrease in concentration *Submicroscopic* illustration of changes in the number of particles when there is an addition/reduction of substances *symbolic*, the shift in the direction of the equilibrium when the concentration of products/reactants is changed
3.9.2 Analyze the temperature factor that affects the direction of the shift in the equilibrium on	C4	*macroscopic*, the change in the color of the substance when the temperature rises / falls *Submicroscopic*, Particle changes when the temperature increases/decreases *symbolic*, the shift in the direction of equilibrium by rising/falling temperature is seen from the value of the Enthalpy Change (ΔH)
3.9.3 Analyzing the volume and pressure factors that affect the direction of the equilibrium shift	C4	*macroscopic* Changes in volume and pressure on a piston *Submicroscopic*, the change in the number of particles when the volume is increased, and the pressure is reduced, and vice versa *symbolic*, the shift in the direction of equilibrium when the volume is increased and the pressure is decreased, and vice versa
3.9.4 Analyzing the application of chemical equilibrium in the industry	C4	*macroscopic*, Haber-Bosch ammonia synthesis process *Submicroscopic*, an illustration of the molecular and particle changes that occur in the ammonia synthesis process *symbolic*, Equation of the reaction in the synthesis of ammonia

Table 2. Achievement of Understanding Levels of Multi Representation of Chemistry

Level	Information
Level 1, connecting symbolic understanding with macroscopic understanding	Students can answer macroscopic-symbolic questions correctly. Or Students can answer one of the macroscopic, submicroscopic, *or* symbolic questions correctly.
Level 2, understand sub-microscopic and symbolic understanding	Students can answer the symbolic-submicroscopic questions correctly.
Level 3, understand and interpret the notion of macroscopic, submicroscopic, and symbolic with macroscopic.	Students can answer all macroscopic-symbolic-submicroscopic questions correctly. Or Students can answer macroscopic-submicroscopic questions correctly.

2) Validation of the contents of the test instrument that has been designed

The questions used as an evaluation tool must already have valid criteria. One way to find out whether the questions that have been designed are valid or not is to do content validity. Content validity is the validity estimated through testing the feasibility or relevance of the test content through rational analysis through a competent panel or expert judgment [12]. The test instrument was carried out by content validation with five validators consisting of 4 chemistry lecturers at the Padang State University Faculty of Mathematics and Natural Sciences and one teacher at SMAN 3 Padang. This content validation test uses a Guttman scale with a choice of "Yes" or "No" as an aspect of the assessment that will be given. The results of content validation were analyzed using the minifac application so that the suitability of the items with the assessment aspects of the instrument based on the validator's assessment could be seen. This information can be seen in the following table from the results of measurement data analysis by the validator using the Rasch model.

Table 3. Validator Measurement Analysis Results

Level Value	Reliability	exact agreements	Expected agreements
2.81	0.78	89.1%	89.7%

Table 2 states that the strata value score obtained is 2.81, close to 3. It indicates that the data collected is reliable [13]. The reliability value obtained is 0.78, which is included in the sufficient category. The reliability value shows the validator's expertise in assessing the questions. Furthermore, the value of the tester's agreement (exact agreement) is 89.1%, while when compared with the results of the estimated agreement (expected agreement), of 89.7% [14]. It shows that the results of the validator's assessment with the expected results are not much different, so the items that have been designed can be said to be valid and have a good category based on the results of the validator's assessment so that it can be concluded that the questions designed can be tested on subjects.

3) Repair of instrument test

The test instrument that has been validated by the validator, apart from containing raw data that can be processed, the results of the questionnaire on content validity also include some suggestions and input from the validator to be taken into consideration in making improvements to the test items so that the test instrument is more feasible to try out.

The following questions are included as examples to provide context for the interconnection of multi-representation questions that are feasible to try out.

Test Question Number 3: Students carry out an experiment using a solution of $Fe(SCN)^{2+}$, which is in beaker A with a red color. The $Fe(SCN)^{2+}$ solution was produced by reacting the orange-colored Fe^{3+} solution with the colorless SCN^- solution.

$$Fe^{3+}(aq) + SCN^-(aq) \leftrightharpoons Fe(SCN)^{2+}(aq)$$

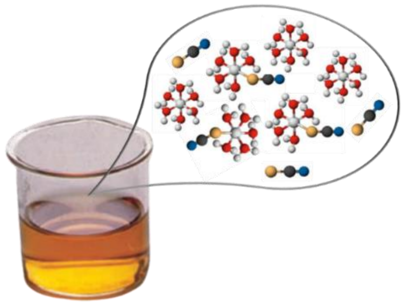

Figure 1. Effect of a change in concentration on a shift in the direction of the equilibrium (Chang R, 2011:527)

From the information given above, define:

a. The reaction for a new equilibrium when glass A is added to a solution of $H_2C_2O_4$ (can bind ions Fe^{3+}) and when glass A is added to a solution of Na(SCN)

b. Look at the pictures of the two beakers below.

Red (1)

Pale Yellow (2)

Determine which color of the solution in the beaker will be produced when the equilibrium system in beaker A is changed, along with the reasons

- State of the system when $H_2C_2O_4$ solution is added (can bind strongly to Fe^{3+})
- State of the system when Na(SCN) solution is added

c. Make an illustration of the particles according to the table below based on the particles present in the equilibrium system in beaker one and beaker two after adding substances

Doing Trials

At this stage, trials were carried out with predetermined subjects. The subjects used were 30 class XII students at SMAN 3 Padang who had studied chemical equilibrium material. The number of subjects selected met the minimum requirements for conducting trials on the Rasch modeling, namely 30 subjects had the same level of confidence as 100 subjects, namely 95%. Before conducting trials with test questions, students remember and are taught the chemical equilibrium material again to avoid biased values. The results are corrected from the students' answers, and then an analysis will be carried out with the ministep application.

Analyzing Data Using Rasch modeling

At this stage, the raw data from students' answers are processed using the Rasch model using the ministep application. The analysis was carried out to review the item questions that had been made, namely: (1) validity, (2) reliability, (3) index of difficulty, (4) discriminatory power (5) Differential Item Functioning. The results of the analysis obtained include the following:

1. Validity

In the Rasch model, the validity test is known as unidimensionality items, namely items used to evaluate whether the instrument can measure what it should measure. One way to look for unidimensionality/item fit items is to measure based on three criteria: outfit Mean Square (MNSQ), outfit Z-Standard (ZSTD), Pt Measure Correlation (Pt Mean Corr) [15]. Each question item must meet two of the three existing criteria to be valid [9].

Item STATISTICS: MISFIT ORDER

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
4	27	30	-1.14	.63	1.19	.55	1.55	.95	A-.09	.23	90.0	90.0	2b	A
7	36	30	.93	.25	1.47	2.07	1.49	1.84	B .21	.49	23.3	45.2	3c	B
10	45	29	.16	.30	1.39	1.32	1.35	.95	C .45	.46	69.0	65.7	4c	B
3	56	30	-1.11	.49	1.36	.77	1.31	.63	D .19	.29	86.7	87.6	2a	B
2	26	30	-.79	.56	1.01	.15	1.21	.56	E .20	.26	86.7	86.7	1b	A
1	56	30	-1.11	.49	.92	.02	1.15	.44	F .11	.29	83.3	87.6	1a	B
16	29	23	.95	.29	1.11	.54	1.11	.48	G .32	.47	39.1	46.2	6c	B
13	38	27	.49	.29	1.04	.25	.96	-.01	H .51	.50	70.4	56.3	5c	B
6	66	30	1.15	.16	1.00	.10	.92	-.20	h .65	.65	26.7	30.6	3b	C
9	80	28	.59	.18	.98	.01	.94	-.05	g .65	.66	25.0	28.7	4b	C
14	45	29	.16	.30	.92	-.18	.82	-.36	f .50	.46	69.0	65.7	6a	B
15	43	28	.18	.31	.85	-.45	.63	-.98	e .77	.47	67.9	63.6	6b	B
8	49	29	-.26	.35	.79	-.54	.84	-.17	d .45	.42	79.3	72.6	4a	B
5	54	30	-.71	.41	.80	-.34	.53	-.72	c .56	.35	90.0	80.1	3a	B
12	38	30	.80	.26	.75	-1.22	.74	-1.01	b .53	.49	53.3	46.4	5b	B
11	51	30	-.29	.35	.40	-2.18	.37	-1.56	a .70	.40	83.3	75.5	5a	B
MEAN	46.2	28.9	.00	.35	1.00	.1	1.00	.1			65.2	64.3		
P.SD	14.0	1.8	.77	.13	.26	.9	.33	.8			23.4	19.7		

Figure 1. Validity Results with Fit Order Items

Table 4. Results of Valid Question Item Analysis

No.	Items	Outfit MNSQ	Outfit ZSTD	PT Mean Corr	Info
1	2b	1.55	0.95	-0.09	Valid
2	3c	1.49	1.84	0.21	Valid
3	4c	1.35	0.95	0.45	Valid
4	2a	1.31	0.63	0.19	Valid
5	1b	1.21	0.56	0.20	Valid
6	1a	1.15	0.44	0.11	Valid
7	6c	1.11	0.48	0.32	Valid
8	5c	0.96	-0.01	0.51	Valid
9	3b	0.92	-0.20	0.65	Valid
10	4b	0.94	-0.05	0.65	Valid
11	6a	0.82	-0.36	0.50	Valid
12	6b	0.63	-0.98	0.77	Valid
13	4a	0.84	-0.17	0.45	Valid
14	3a	0.53	-0.72	0.56	Valid
15	5b	0.74	-1.01	0.53	Valid
16	5a	0.37	-1.56	0.70	Valid

In this analysis, it was found that from the sixteen sub-questions, two of the three criteria were met. In the Pt Measure Corr criterion, most were not met because a value <0.4 produced six items which can be seen in table 4 for item questions 2b,3c,2a,1b,1a,6c. In the MNSQ criteria with a value limit of 0.5 < MNSQ < 1.5, one item is not fulfilled, namely in question 5a. However, the seven items are still eligible to be maintained without revision and declared valid because they meet the other two criteria. In addition, all item items successfully meet the score criteria with a range of -2.0 < ZSTD < +2.0 because this ZSTD acts as a t-test in the data suitability hypothesis, which means that all item items have logical data estimates. From the item fit analysis results, it can be concluded that all questions meet 2 of the three criteria, so all item items are said to be valid.

2. Reliability

Reliability results were obtained through analysis of the Output table—summary Statistics. Two criteria can be seen to determine which items are reliable, namely by reviewing Cronbach's Alpha scores and item reliability.

CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .74 SEM = 2.88
STANDARDIZED (50 ITEM) RELIABILITY = .86

SUMMARY OF 16 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	46.2	28.9	.00	.35	1.00	.06	1.00	.05
SEM	3.6	.5	.20	.03	.07	.24	.08	.22
P.SD	14.0	1.8	.77	.13	.26	.94	.33	.85
S.SD	14.5	1.8	.79	.13	.27	.97	.34	.87
MAX.	80.0	30.0	1.15	.63	1.47	2.07	1.55	1.84
MIN.	26.0	23.0	-1.14	.16	.40	-2.18	.37	-1.56

REAL RMSE .39 TRUE SD .66 SEPARATION 1.68 Item RELIABILITY .74
MODEL RMSE .37 TRUE SD .67 SEPARATION 1.80 Item RELIABILITY .76
S.E. OF Item MEAN = .20

Figure 2. Reliability Results with Summary Statistics

Cronbach's Alpha value is used to measure reliability, namely the interaction between the subject and the question items as a whole [15]. In the statistical summary table, it can be seen that Cronbach's Alpha value is 0.74 and is included in the good category. The reliability item value is 0.74 and is included in the sufficient category. Based on the analysis results on Cronbach's Alpha and the reliability items that match the criteria, the item questions on the test instrument are reliable.

3. Difficulty Index

The difficulty index can be analyzed through the Output Table menu. Item Measures. Through Item measure, the order of difficulty level from the most difficult to the easiest based on logit can be seen in the measure and item sections. The results of the analysis of the difficulty index of this instrument are as follows:

Item STATISTICS: MEASURE ORDER

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
6	66	30	1.15	.16	1.00	.10	.92	-.20	.65	.65	26.7	30.6	3b	C
16	29	23	.95	.29	1.11	.54	1.11	.48	.32	.47	39.1	46.2	6c	B
7	36	30	.93	.25	1.47	2.07	1.49	1.84	.21	.49	23.3	45.2	3c	B
12	38	30	.80	.26	.75	-1.22	.74	-1.01	.53	.49	53.3	46.4	5b	B
9	80	28	.59	.18	.98	.01	.94	-.05	.65	.66	25.0	28.7	4b	C
13	38	27	.49	.29	1.04	.25	.96	-.01	.51	.50	70.4	56.3	5c	B
15	43	28	.18	.31	.85	-.45	.63	-.98	.77	.47	67.9	63.6	6b	B
14	45	29	.16	.30	.92	-.18	.82	-.36	.50	.46	69.0	65.7	6a	B
10	45	29	.16	.30	1.39	1.32	1.35	.95	.45	.46	69.0	65.7	4c	B
8	49	29	-.26	.35	.79	-.54	.84	-.17	.45	.42	79.3	72.6	4a	B
11	51	30	-.29	.35	.40	-2.18	.37	-1.56	.70	.40	83.3	75.5	5a	B
5	54	30	-.71	.41	.80	-.34	.53	-.72	.56	.35	90.0	80.1	3a	B
2	26	30	-.79	.56	1.01	.15	1.21	.56	.20	.26	86.7	86.7	1b	A
1	56	30	-1.11	.49	.92	.02	1.15	.44	.11	.29	83.3	87.6	1a	B
3	56	30	-1.11	.49	1.36	.77	1.31	.63	.19	.29	86.7	87.6	2a	B
4	27	30	-1.14	.63	1.19	.55	1.55	.95	-.09	.23	90.0	90.0	2b	A
MEAN	46.2	28.9	.00	.35	1.00	.1	1.00	.1			65.2	64.3		
P.SD	14.0	1.8	.77	.13	.26	.9	.33	.8			23.4	19.7		

Figure 3. Difficulty Index Test Results with Item Measure

The main criterion in classifying the difficulty index instruments in Figure 3 is the standard deviation (SD). This instrument has an SD of 0.77. Therefore, the problem difficulty groups are obtained as follows:

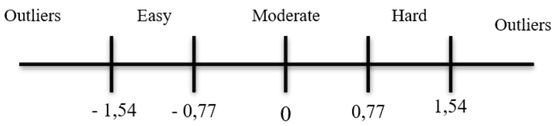

Figure 4. Problem Difficulty Level Groups

Based on the grouping of the difficulty level of the questions above, items with a logit between -0.77 to 0.77 are questions with medium difficulty. Items with a logit between -0.77 to -1.54 are easy questions. Items with a logit between 0.77 to 1.54 are difficult questions. While items with logit outside these values are outlier items (too easy/too difficult).

Comparison of the percentage of difficulty in this instrument, namely: difficult (25%), moderate (50%), and easy (25%) (table 5). The results of the analysis prove that this test instrument has a good index of difficulty because, overall, the items are relatively easy and easy [16].

Table 5. Results of the Problem Index Difficulty Analysis

Category	Items	Login	Percentage
Hard	3b	1.15	25%
	6c	0.95	
	3c	0.93	
	5b	0.80	
Moderate	4b	0.59	50%
	5c	0.49	
	6b	0.18	
	6a	0.16	
	4c	0.16	
	4a	-0.26	
	5a	-0.29	
	3a	-0.71	
Easy	1b	-0.79	25%
	1a	-1.11	
	2a	-1.11	
	2b	-1.14	

4. Difference Power

Discriminating power, or item discrimination, can be analyzed using the Output table. Summary Statistics. The results of this analysis can be seen in Figure 4. The separation results of 1.68 are then processed to see a more thorough grouping called strata separation using the following H formula.

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

$$H=\frac{[(4\times 1{,}68)+1]}{3}$$

$$H=2{,}57$$

The result of the separation point value is 2.57, then it is rounded up to 3, meaning that there are three groups of item items: difficult, medium, and easy.

5. Differential Item Functioning

Detection of bias in the item items in the Rasch model analysis is shown in the differential item functioning (Differential Item Functioning or DIF). This analysis is needed to determine which items are biased in certain categories of respondents or not [17]. DIF is defined as the conditional probability of obtaining the correct answer in the form of items that are assessed dichotomously by people with the same abilities but with different demographic features [18]. DIF refers to the idea that gender, ethnicity, or age groups respond differently, even though they share the same degree of the latent trait. Bias towards gender may be a possible reason for misleading interpretations of gender differences in a given construct or may show incorrect differences. DIF detection is an important test construction and validation procedure [19].

```
    SUMMARY OF 16 MEASURED Item
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.     MNSQ   ZSTD   MNSQ   ZSTD  |
|-----------------------------------------------------------------------------|
| MEAN      46.2      28.9         .00     .35     1.00    .06   1.00    .05  |
|  SEM       3.6        .5         .20     .03      .07    .24    .08    .22  |
| P.SD      14.0       1.8         .77     .13      .26    .94    .33    .85  |
| S.SD      14.5       1.8         .79     .13      .27    .97    .34    .87  |
| MAX.      80.0      30.0        1.15     .63     1.47   2.07   1.55   1.84  |
| MIN.      26.0      23.0       -1.14     .16      .40  -2.18    .37  -1.56  |
|-----------------------------------------------------------------------------|
| REAL RMSE    .39 TRUE SD     .66  SEPARATION  1.68 Item   RELIABILITY  .74 |
|MODEL RMSE    .37 TRUE SD     .67  SEPARATION  1.80 Item   RELIABILITY  .76 |
| S.E. OF Item MEAN = .20                                                     |
-------------------------------------------------------------------------------
```

Figure 4. The results of the Power Test are different from the Summary Statistics

Based on the probability value in Figure 5, it can be seen that there are no item items that are biased for gender criteria. All probability scores for the sixteenth sub-questions are above 5% (0.05). It shows that the question item does not harm certain gender groups.

J. Pijar MIPA, Vol. 18 No. 2, March 2023: 208-218
DOI: 10.29303/jpm.v18i2.4720
ISSN 1907-1744 (Cetak)
ISSN 2460-1500 (Online)

DIF class/group specification is: DIF=$S3W1

Person CLASSES	SUMMARY DIF CHI-SQUARED	D.F.	PROB.	BETWEEN-CLASS/GROUP UNWTD MNSQ	ZSTD	Item Number	Name
2	.0126	1	.9107	.0119	-1.17	1	1a
2	1.5436	1	.2141	1.7835	.92	2	1b
2	.0126	1	.9107	.0119	-1.17	3	2a
2	2.0241	1	.1548	3.1823	1.47	4	2b
2	.4834	1	.4869	.5205	.06	5	3a
2	.0079	1	.9294	.0106	-1.18	6	3b
2	3.3242	1	.0683	4.2392	1.78	7	3c
2	.8147	1	.3667	.8900	.39	8	4a
2	.5052	1	.4772	.5452	.08	9	4b
2	.2706	1	.6029	.2944	-.24	10	4c
2	.1094	1	.7408	.1159	-.62	11	5a
2	2.5204	1	.1124	3.0569	1.43	12	5b
2	.0477	1	.8271	.0492	-.87	13	5c
2	1.5409	1	.2145	1.7422	.90	14	6a
2	1.4003	1	.2367	1.5867	.82	15	6b
2	.3676	1	.5443	.4236	-.06	16	6c

Figure 5. Probability Value of Measuring DIF

The size DIF curve is a curve that shows students' responses to the test instrument based on gender. The curve shows the range of sex differences in answering the question items indicated by the normal green line [20]. The further the range of male (black) and female (red) answers, the greater the bias of the questions in distinguishing gender characteristics. Goods received are right on or around the normal line [9].

In addition to the number of items that have a bias (DIF) which can be described in graphical form, it can be seen from the graph that the curve that approaches the upper limit is a question with a high degree of difficulty in items 3b, 3c, 5b, 6c. It can be seen that the items are easier for female students to work on than male students, while the curve below shows the items that are easy to work on.

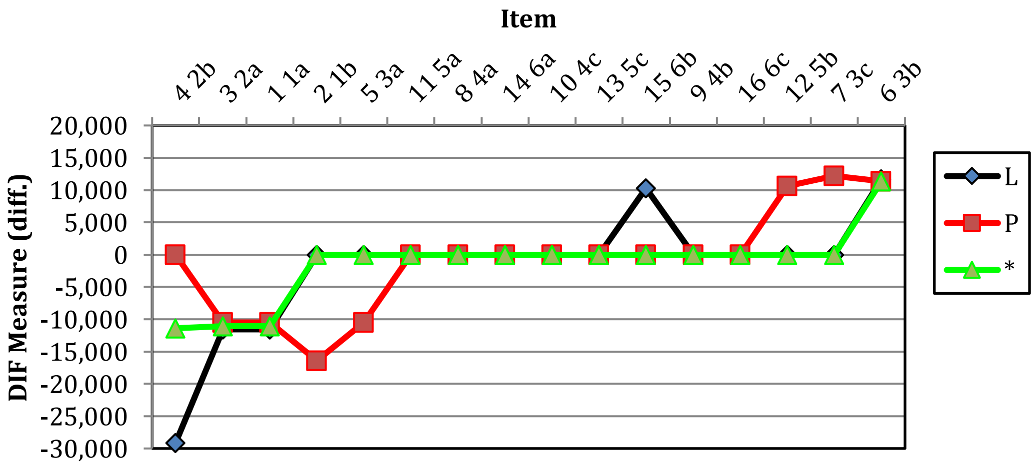

Figure 6. DIF Curve Person DIF plot (DIF=$S3W1)

Review Item Compatibility

The analysis review using the Rasch modeling shows that the results of the validity, reliability, difficulty index, and differentiability of all item items on the test instrument have fulfilled the criteria or are following the Rasch modeling. The instrument's validity includes valid, good instrument reliability and sufficient item reliability, the overall item difficulty index is balanced, and the differential power of the instrument is good.

Viewing the Wright Map

Analysis on the Wright map can reach all levels of student abilities (high to low) and the difficulty level of the item questions (easy to difficult). From the image of the wright map, it can be seen that there are two side panels, namely the left and right. The left panel on the Wright map shows students' ability level to answer the item questions, while the right panel shows the questions' difficulty level. Wright's map states that the higher the cognitive processes students have, the quality of the item question must have a higher difficulty level. It is because the quality of the items is directly proportional to the student's ability level [21-22]. The results of the analysis of the wright map in the picture can be seen on the right panel of students with codes 14 NA, 15SR,

J. Pijar MIPA, Vol. 18 No. 2, March 2023: 208-218 ISSN 1907-1744 (Cetak)
DOI: 10.29303/jpm.v18i2.4720 ISSN 2460-1500 (Online)

26HZ, students with higher abilities than others with a logit value of +2.44. This indicates a high ability of students to answer the whole question items, including difficult questions on the right panel of item number 3b.

Students with code 03RG have the lowest ability to work on question items with a logit value of -0.41 seen on the right of the item distribution panel located in the easy item section, namely in question codes 1a,1b,2a,2b,3a,4a, and 5a. The item questions are being spread evenly at a logit of 0.0 to +1 for eight items. If we look at the standard deviation (SD) on the Wright map, the distribution in the left panel is for students' abilities, and the right panel is for item items with a range of +2SD and -2SD, so the data is spread up to 95%.[9]. From the results of the distribution of all item items, if you look at the distribution of the existing logit scale, it is not outside the T (outlier) limit, so the item items do not require revision and can be maintained. Based on a review of the Wright map, it was found that the test instrument had eligibility because the items were spread out in the medium category, so it was neither too easy nor too difficult.

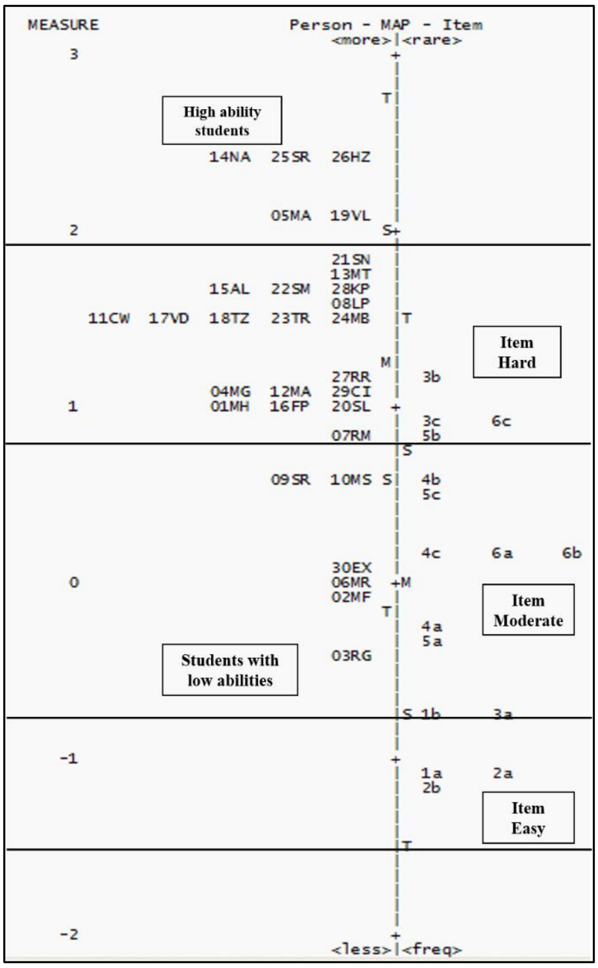

Figure 5. Results of the Wright Map Review

After the item is declared valid, the resulting test instrument can also measure the level of students' understanding of the chemical equilibrium material as a whole. The achievement criteria at each level of understanding can be seen in Table 2. It can be analyzed from students' answers to the item questions to determine their achievements at each level. For example, in question number 3, with a key GPA of 3.9.1, namely analyzing concentration factors that affect the direction of shifting the equilibrium, question number 3 has sub-items that represent the three levels of representation (sub-item a is a symbolic representation, sub-item b is macroscopic representation, sub-item c is representation symbolic). In question number 3 for the answer to item number 3a, which represents a symbolic representation, namely being able to develop a reaction based on the phenomenon described complete with the phrase, then in answer 3b, which represents a macroscopic representation, namely being able to determine the color of the solution produced when the concentration in the system is changed, being able to determine the direction the shift in equilibrium when the concentration changes and being able to explain down to the particulate level can be seen from answer 3c which can illustrate the illustration of the particles that occur when the concentration in the system is changed. It can be concluded that students are at level 3 because they can make the reaction equations correctly; state and explain macroscopic phenomena and processes from sub-microscopic to problem numbers.

Level 3 achievement can be achieved if students answer all the sub-items representing each level of representation. Participants' answers were at level 3 for question number 3 if they could answer questions from all three levels of representation (3a a maximum score of 2, 3b a maximum score of 4, 3c a maximum score of 2). It means that students have level 3 understanding because they can answer questions at the macroscopic (3b), sub-microscopic (3c), and symbolic (3a) understanding levels, or students have level 3 achievement if they are only able to answer questions between macroscopic level sub-representations (3b) with sub-microscopic (3c) even though the question for the symbolic level (3a) is wrong.

In question number 3, for answers that get achievements at level 2, that is, if sub-item 3b is answered incorrectly (score 0), the answer for the macroscopic representation sub-item is not answered correctly. In symbolic representation, students can make reactions based on the phenomena described complete with their phases, then in answer 3b which represents macroscopic representations students do not get the maximum score because students are not able to explain how the direction of the shift in equilibrium can change when the concentration on the system is changed, to get a maximum score of 4 on point 3 students must answer with the range of answers for the state of the system when added to the $H_2C_2O_4$ solution the resulting color in glass 2 is pale yellow, and the equilibrium shifts from left to right because oxalate ions bind strongly with Fe^{3+} this will make Fe^{3+} is lost from the solution and more $FeSCN^-$ is produced and when the system is added to the Na(SCN) solution the color produced in glass 1 is red and the equilibrium shifts from right to left due to an increase in SCN^- concentration, so Fe^{3+} ions react to the pen the addition of SCN^- ions so that more $Fe(SCN)^{2+}$ is formed causing the dominant color to red. Whereas for answer 3c the illustration of the particle which is described exactly corresponds to the situation that occurs when $H_2C_2O_4$ is added to the system, more reactants are formed which are characterized by more rectangular and circular images, whereas when Na(SCN) solution is added, the number of images is greater. A triangle which is the product, the red color, is produced due to the increase in SCN^- concentration, so Fe^{3+} ions react to the addition of SCN^- ions so that more $Fe(SCN)^{2+}$ is red.

In question number 3 an answer with a level 1 representation achievement is if the answer to sub-item 3c is answered incorrectly (score 0) while the maximum score for 3c is 2, which means the answer for the sub-item sub-microscopic representation was not answered correctly. It can be seen from answer 3a, which represents a symbolic representation; students can make reactions based on the phenomenon described complete with its phases, then in answer 3b, which represents a macroscopic representation. Because the answer is correct, it is included in achieving level 1 representation, namely, connecting the symbolic and macroscopic levels.

Based on the results of the Wright map, it can be seen that in the difficult part of the questions, there are questions with numbers 3b, 6c, 3c, and 5b, which have a distribution of sub-items at the macroscopic and sub-microscopic levels, for medium questions, there are three distributions for the representation level, namely questions at the macroscopic level (numbers 4b and 6b), a symbolic level (numbers 5c and 4c) and macroscopic level questions (numbers 6a, 4a, 5a, 3a). At the same time, the questions in the easy category were dominated by the symbolic level (numbers 1a, 1b, 2a, and 2b). From this description, questions with the symbolic level dominate the questions with the easy category. While questions with categories that are difficult to reach by students are questions with macroscopic representation with a total logit of +0.97 and sub-microscopic with a logit of +0.94 so that it can be interpreted that the learning process of students requires learning that links chemical equilibrium material related to macroscopic and sub-microscopic.

Repeating Steps 4-7 Until All Items Fit

At this stage, no repetition was carried out in steps 4-7 because, based on the results of the data that has been analyzed, it fits with the criteria of the Rasch modeling.

Establish Claims

At this stage, it was determined that based on the overall analysis of the item questions on the test instrument, it could be used to evaluate the concept of chemical equilibrium at the macroscopic, submicroscopic, and symbolic levels of students because they had been tested for quality in terms of validity, reliability, difficulty index, and discriminatory power.

Documenting Instruments

The last stage is carried out to provide information related to the test instrument that has been developed. It will be useful in properly applying the test instrument to its users. Important information included in this documentation is the purpose of using test instruments, learning progression, question indicators, item items, assessment rubrics, and guidelines for analyzing students' understanding levels related to macroscopic, sub-microscopic, and symbolic levels in chemical equilibrium material.

CONCLUSION

The test instrument developed to evaluate the concept of chemical equilibrium at the macroscopic, sub-microscopic, and symbolic levels of students in chemical equilibrium material has been tested for valid, reliable quality and has a good index of difficulty and discriminatory power. The quality of this test instrument can be seen from the results of the research that there is a suitability of validation by the validator using the minifacets program with the acquisition of exact agreements and expected agreements not much different, namely 89.1% and 89.7%, this indicates that this analysis is in accordance between the models and estimates. Likewise, in analyzing raw data obtained from students after being processed using the ministep program, each item's results can be considered valid because they meet the criteria of MNSQ, ZSTD, and PtMean Corr. The test instrument is also reliable in the good category when viewed from the Cronbach Alpha value. The items are reliable in the sufficient category because it has a value of 0.74. The difficulty index is in a good category because the questions are spread out in moderate groups so that they are easy and manageable and have a good distinction in the easy, medium, and difficult categories. All question items also do not have DIF based on gender.

REFERENCES

[1] Stojanovska, M., Petruševski, V. M., & 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). doi: 10.20903/csnmbs.masa.2014.35.1.52.

[2] Akaygun, S. (2016). Is the oxygen atom static or dynamic? The effect of generating animations on students' mental models of atomic structure. Chemistry Education Research and Practice, 17(4), 788-807.

[3] Harza, A. E. K. P., Wiji, W., & Mulyani, S. (2021, March). Potency to overcome misconceptions by using multiple representations on the concept of chemical equilibrium. In Journal of Physics: Conference Series (Vol. 1806, No. 1, p. 012197). IOP Publishing.

[4] Sunyono, S., Yuanita, L., & Ibrahim, M. (2015). Mental models of students on stoichiometry concept in learning by method based on multiple representation. The Online Journal of New Horizons in Education (TOJNED), 5(2), 30-45.

[5] Baig, M., Ali, S. K., Ali, S., & Huda, N. (2014). Evaluation of multiple choice and short essay question items in basic medical sciences. Pakistan journal of medical sciences, 30(1), 3.

[6] Aristiawan, A. (2022). Profil Kemampuan Pemecahan Masalah Fisika Siswa SMA Menggunakan Soal Essay. Jurnal Tadris IPA Indonesia, 2(1), 45-55.

[7] Son, A. L. (2019). Instrumentasi kemampuan pemecahan masalah matematis: analisis reliabilitas, validitas, tingkat kesukaran dan daya beda butir soal. Gema wiralodra, 10(1), 41-52.

[8] Stolt, M., Kottorp, A., & Suhonen, R. (2022). The use and quality of reporting of Rasch analysis in nursing research: A methodological scoping review. International Journal of Nursing Studies, 104244.

[9] Sumintono, B., & Widhiarso, W. (2015). Aplikasi pemodelan rasch pada assessment pendidikan. Trim komunikata.

[10] 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.

[11] 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.

[12] Utomo, B. (2022). Analisis validitas isi butir soal sebagai salah satu upaya peningkatan kualitas pembelajaran di madrasah berbasis nilai-nilai islam. Jurnal Pendidikan Matematika (Kudus), 1(2).

[13] Fahmina, S. S., Masykuri, M., Ramadhani, D. G., & Yamtinah, S. (2019, December). Content validity uses Rasch model on computerized testlet instrument to measure chemical literacy capabilities. In AIP Conference Proceedings (Vol. 2194, No. 1, p. 020023). AIP Publishing LLC.

[14] Nisa, D. Q., & Yusmaita, E. (2022). Development of Chemical Literacy Items in Electrolyte and Non Electrolyte Solution Topic for Class X SMA/MA. Entalpi Pendidikan Kimia, 3(1), 49-57.

[15] Rahman, Y. A. (2023). Person and Item Validity and Reliability in Essay Writing Using Rasch Model. Konstruktivisme: Jurnal Pendidikan dan Pembelajaran, 15(1), 41-55.

[16] Palimbong, J., Mujasam, M., & Allo, A. Y. T. (2018). Item Analysis Using Rasch Model in Semester Final Exam Evaluation Study Subject in Physics Class X TKJ SMK Negeri 2 Manokwari. Kasuari: Physics Education Journal (KPEJ), 1(1), 43-51.

[17] Sumintono, B., & Widhiarso, W. (2014). Aplikasi model Rasch untuk penelitian ilmu-ilmu sosial (edisi revisi). Trim Komunikata Publishing House.

[18] LA, N. (2021). Differential Item Functioning and Test Performance: a Comparison Between the Rasch Model, Logistic Regression and Mantel-Haenszel.

[19] Khalaf, M. A., & Omara, E. M. N. (2022). Rasch analysis and differential item functioning of English language anxiety scale (ELAS) across sex in Egyptian context. BMC psychology, 10(1), 242.

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

[21] Kristyasari, M. L., Asmaningrum, H. P., & Hanif, R. F. Wright-Map to Analyze Students' Abilities on Chemical Bond Test. In SHS Web of Conferences (Vol. 149, p. 01050). EDP Sciences.