JPSI 9(3):329-345, 2021
Jurnal Pendidikan Sains Indonesia
e-ISSN: 2615-840X p-ISSN: 2338-4379

Analysis of Rasch Model for the Validation of Chemistry National Exam Instruments

Ayi Darmana[1], Ani Sutiani[2], Haqqi Annazili Nasution[2], Ismanisa[3*], Nurhaswinda[4]

[1,2]Study Program FMIPA of Chemistry, State University of Medan, Medan, Indonesia
[3]Study Program Postgraduate of Chemistry Education, State University of Medan, Medan, Indonesia
[4]Study Program FKIP of Primary Education, University Pahlawan Tuanku Tambusai, Riau, Indonesia.

*Email: iismanisa21@gmail.com

DOI: 10.24815/jpsi.v9i3.19618

Article History: *Received: January 21, 2021* *Revised: March 21, 2021*
Accepted: April 30, 2021 *Published: May 24, 2021*

Abstract. Information about score obtained from a test is often interpreted as an indicator of the student's ability level. This is one of the weaknesses of classical analysis that are unable to provide meaningful and fair information. The acquisition of the same score if it comes from a test item with a different level of difficulty, must show different abilities. Analysis of the Rasch model will overcome this weakness. The purpose of this study was to analyze the quality of the items by validating the national chemistry exam instrument using the Rasch model. The research sample was 212 new students of the Department of Chemistry at the State University of Medan. The data collected was in the form of respondent's answer data to the 2013 chemistry UN questions, which amounted to 40 items multiple choice and uses the documentation method. Data analysis technique used the Rasch Model with Ministep software. The results of the analysis show the quality of the Chemistry National Exam (UN) questions is categorized as very good based on the following aspects: unidimension, item fit test, person map item, difficulty test level, person and item reliability. There is one item found to be gender bias, in which men benefit more than women. The average chemistry ability of respondents is above the average level of difficulty of the test items.
Keywords: National exam, Dichotomy, Rasch model, Ministep

Introduction

The test is one way to measure the level of human ability indirectly, namely through a person's response to a number of stimuli or questions (Mardapi, 2008). A good quality test has the characteristics of a good test item and a test kit that is known through measurement. Measurement is the process of giving numbers which are expected to show the ability of students about a subject. Measurement is one of the first steps in the evaluation program, which is a process to determine the characteristics of a number of attributes of students, especially the abilities of students (Susongko, 2016). To take measurements, a measuring instrument is needed that provides information about a person's position in the measured characteristics. A good measuring tool will ensure valid and reliable results so that it can measure students' abilities accurately.

At this time, many educators still use the classical theory measurement model, even though this classical theory has several weaknesses. An alternative model to overcome the weaknesses in classical theory is to use the Rasch model measurement (Chan, et al., 2013).

The Rasch model uses one parameter in analyzing the test taker's ability, with the application used is the Ministep Software. Analysis with the Rasch model is quite easy to do but produces accurate analysis results. Rasch reviewed the chance of answering correctly on dichotomous form questions by comparing students' abilities with the difficulty level of the questions. Thus, students have a 50% chance of answering the questions correctly, if it is known that the students' abilities are the same as the difficulty level of the questions. This is in accordance with the opinion expressed by Sumintono & Widhiarso (2015) and Linacre (2016) that the Rasch model has several advantages, namely the Rasch model can identify error responses, predict missing data scores, differentiate the ability of respondents with the same raw score, can analyze data dichotomy and polytomy and their combination, and identify indications of guesswork and cheating.

Previous research using the Rasch model as a computer program in testing measuring instruments has been carried out, including in the measurement of test instrument questions by Ibrahim, et al., 2015; Mahmud, et al., 2017; Wati, et al., 2019; Sihombing, et al., 2019; Isnani, et al., 2019; Samritin, et al., 2019; Saidi & Siew, 2019; Pratama & Husnayaini, 2020; Darmana & Sutiani, 2020. Rasch model can produce standard error measurement values that can improve calculation accuracy (Afrassa, 2005; Ardiyanti, 2016). Rasch model is recommended for use in test instrument analysis (Sabekti & Khoirunisa, 2018). Based on this explanation, the Rasch model is an assessment analysis model that is recommended to be used by educators in measuring and assessing student learning outcomes to determine the true abilities of students. Based on these reasons, this study will analyze the quality of the National Examination test instrument and the initial abilities of chemistry new students using the Rasch model.

Methods

This research is descriptive research. The study was conducted in early September 2020. The sample of this study was 212 new students of the Chemistry Department at Medan State University in the academic year 2020/2021. The test instrument used was the 2013 Chemistry national examination (UN) question instrument consisting of 40 multiple choice questions (dichotomy) with five response categories. Based on the test instrument, the results of the test takers' answers were obtained and collected through the documentation method. The data analysis technique used the Rasch model with the help of WINSTEPS version 3.73 and SPSS version 19.0 software. The Rasch model was chosen because the Rasch model can review the chance of answering correctly on dichotomous questions by comparing students' abilities with the difficulty level of the questions. The instrument validation aspects analyzed included the Rasch model prerequisite test, namely the local unidimensional test and independence (Bond & Fox, 2007), item fit, item difficulty and person ability (wright map), bias test with the DIF (differential item functioning), reliability, and the measurement information function. Criteria for a valid test viewed from various aspects and criteria can be seen in Table 1.

Table 1. Criteria for a valid test viewed from various aspects and criteria

The Validity Aspect of the Item	Criteria
Unidimension test	There is one main factor that is pictures through Screen Plot's factor analysis result
Local independent	The variance-covariance matrix close to 0.00.
Fit item test	0.5 < MNSQ < 1.5 -2.0 < ZSTD<2.0 0.4 < Pt Measure Corr < 0.85
Item difficulty	Very difficult: b (measuring item) > 1; Difficult: 0.5 ≤ b <1; Moderate: -0.5 ≤ b <0.5; Easy: -0.5 ≤ b <-1; and Very easy: b ≤ -1.
Person ability (wright map)	All level of item difficulties are in the testee's domain capability
DIF	Significant DIF found
Reliability Person/Item	Person/Item reliability: Good: 0.81 – 0.90 Very good: 0.91 – 0.94 Special: > 0.94
Alpha Cronbach	Alpha Cronbach > 0.8 is good category
Measurement information functions	Information function test has maximal values on the testee's domain capability

Results and Discussion

Rasch analysis is a mathematical modeling approach based on latent properties and achieves additivity of sticky conjoin (probability), conjoin means measuring persons and items on the same scale (Bond & Fox, 2015). The aim of Rasch's analysis is to maximize trait homogeneity and to allow for greater redundancy without reducing measurement information by item or rating level to produce a more valid and simpler measure. The basic requirements for the Rasch model that need to be considered are unidimensional, item fit, difficulty / ability estimation, reliability, and measurement information functions.

Unidimensional Test. Unidimensional aims to test each instrument item that can measure one variable or only one ability (Reckase, 1979; Susongko, 2016). Unidimensional is also known as the construct validity of an instrument. Factor analysis was used to obtain the dimensions of the instrument. The purpose of factor analysis is to identify the relationship between variables by looking for computational results on the Eigenvalues in the variance-covariance matrix.

The unidimensional assumption test is carried out based on the Kaiser-Meyer-Olkin (KMO) analysis and the Bartlett Sphericity test to determine whether the data obtained is in accordance with factor analysis or not. A measure of the adequacy of sampling or whether the data can be factored well, the KMO-MSA value is greater than 0.6 and the Bartlett Sphericity test must be significant at α <0.05 (Kaiser, 1974; Field, 2009). The results of the KMO-MSA and Bartlett Sphericity tests can be seen in Table 2.

Table 2. The result of KMO-MSA and bartlett sphericity test

Test	Initial ability	Factor analysis result
KMO-MSA *test*	0,808	suitable
The significance value of Bartlett Sphericity *test**	0,00	suitable

As shown in Table 2, the resulting KMO value is greater than 0.6 (0.808> 0.6). The Bartlett Sphericity test shows that the α value is less than 0.05 (0.00 <0.05). Based on the analysis, it can be concluded that the data obtained in this study are suitable for unidimensional factor analysis or construct validity. Construct validity aims to determine whether the instrument item is valid or not in accordance with empirical data. The construct validity was carried out by interpreting the anti-image values obtained after the KMO-MSA and Bartlett Sphericity tests were fulfilled. Factor analysis in proving construct validity with anti-image correlation for all items must be greater than 0.5. The anti-image correlation result has a value greater than 0.5 for each of the 40 items. Thus, the value of this item has a high contribution to the factor structure of the instrument.

Scree plots are another way to define unidimensions. The scree plot is used to illustrate the Eigenvalues with the number of components that can maintain the factor. Unidimensions are declared fulfilled if the instrument has a dominant component that measures the ability being tested (Guler, et al., 2014). If there is a dominant factor with a cumulative percentage greater than 20%, then unidimensions are fulfilled (Barret, et al., 2016). The results of the unidimensional scree plot can be seen in Figure 1.

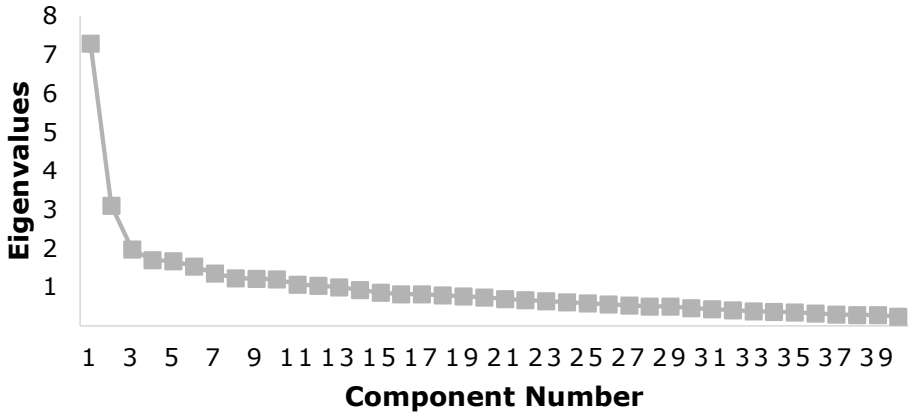

Figure 1. The results of unidimensional scree plots from 40 items.

In Figure 1, it shows that there is only one sharp steepness, namely from component 1 to component 2, while the other components graph looks sloping and does not show a sharp steepness. This explains that the amount of steepness indicates the number of fac-

tors, and other changes in the eigenvalues of the steepness do not indicate a factor (Chernyshenko, et al., 2001). Because there is only 1 steep, namely from component 1 to component 2, while the other steeples cannot be considered as a factor. So this shows that there is only 1 factor being measured (Susongko, 2016), namely the initial chemistry ability of students.

The unidimensional test can also be analyzed using the Winstep program, which can be seen in Figure 2.

	-- Empirical --		Modeled
Total raw variance in observations =	56.1 100.0%		100.0%
Raw variance explained by measures =	16.1 28.6%		28.4%
Raw variance explained by persons =	5.8 10.3%		10.2%
Raw Variance explained by items =	10.3 18.4%		18.2%
Raw unexplained variance (total) =	40.0 71.4%	100.0%	71.6%
Unexplned variance in 1st contrast =	3.7 6.6%	9.3%	
Unexplned variance in 2nd contrast =	3.2 5.7%	8.0%	
Unexplned variance in 3rd contrast =	1.9 3.3%	4.7%	
Unexplned variance in 4th contrast =	1.8 3.2%	4.5%	
Unexplned variance in 5th contrast =	1.7 3.1%	4.3%	

Figure 2. Unidimensional test using Winstep program.

In unidimensional output, the Eigenvalue or raw variance data is obtained by 28.6%. These results indicate that the unidimensionality of the instrument with a minimum value of 20% is fulfilled (Smits, et al., 2011; Wu, et al., 2013; Sumintono & Widhiarso, 2015). The instrument developed can measure what should be measured (Lia, et al., 2020). The unexplained variance was 6.6; 5.7; 3.3; 3.2; and 3.1. This shows that the unexplained variance by the instrument is all less than 10%, which means that the uniformity in the instrument is in the good category (Wibisono, 2019).

Furthermore, local independent assumptions are made, with the aim of proving that participants' responses to one item do not affect responses to other instrument items. Local independence is based on the results of measuring the person output that is sorted from highest to lowest, and then processed by creating a variance-covariance matrix (Greiff, et al., 2013). The local independent assumption is fulfilled when the value below the diagonal line on the variance-covariance matrix approaches 0.00. This value shows that the participants' ability to answer items does not affect their ability to answer other instrument items. Table 3 shows that the covariance value in the initial chemistry ability is close to 0.

Table 3. The result of covariance matrix in the initial chemistry ability

Colums	*K1*	*K2*	*K3*	*K4*	*K5*	*K6*	*K7*	*K8*	*K9*	*K10*
K1	0.287									
K2	0.056	0.013								
K3	0.028	0.007	0.005							
K4	0.011	0.003	0.002	0.003						
K5	0.027	0.006	0.003	0.001	0.004					
K6	0.049	0.011	0.007	0.004	0.006	0.015				
K7	0.047	0.010	0.005	0.003	0.005	0.011	0.010			
K8	0.054	0.012	0.008	0.005	0.006	0.015	0.011	0.017		
K9	0.053	0.013	0.007	0.006	0.005	0.015	0.011	0.017	0.020	
K10	0.184	0.036	0.018	0.000	0.024	0.026	0.028	0.030	0.012	0.160

Based on the Table 3, it shows the results of the variance-covariance value between groups of students' abilities. It can be seen that the covariance value between groups of ability intervals located on the diagonal part is small and close to 0. It can be concluded that there is no correlation, so it can be said that the local independence assumption test is fulfilled.

Item Fit Test. In the Rasch measurement, the concept of fit indicates that the quality of the regulated instruments is adequate. The item fit concept is also used to assess the meaning of the unidimensional construct, which means that the item fit index helps the researcher ensure that Rasch's requirements for the dimension apply empirically. The criteria values used to check the suitability of the items are as follows, (a) the MNSQ value received: 0.5 <MNSQ <1.5; (b) accepted ZSTD value: -2.0 <ZSTD <+2.0; (c) correlation of measurement points (Pt Mean Corr) value: 0.4 <Pt Measure Corr <0.85 (Boone, Staver & Yale, 2014). The results of fit items in measuring students' initial chemistry abilities can be seen in Table 4.

Table 4. Item fit test

Item	*Outfit* MNSQ	*Outfit* ZSTD	*Pt-Measure Correlation*	Category
40	3,08	4,1	-0,18	not suitable
9	1,74	4,9	0,02	not suitable
35	1,40	1,3	0,17	suitable
28	1,37	3,6	0,19	not suitable
27	1,32	3,8	0,15	not suitable
36	1,29	3,5	0,18	not suitable
37	1,26	3,0	0,19	not suitable
14	1,24	1,6	0,44	suitable
24	1,19	1,4	0,33	suitable
5	1,17	2,0	0,28	suitable
29	1,16	2,0	0,25	suitable
25	1,15	1,5	0,29	suitable
16	1,14	1,6	0,31	suitable
15	1,14	1,7	0,30	suitable
38	1,13	1,4	0,32	suitable
33	1,12	1,2	0,28	suitable
12	1,07	0,5	0,28	suitable
6	1,01	0,1	0,36	suitable
2	0,96	-0,3	0,38	suitable
17	1,01	0,1	0,40	suitable
1	0,98	-0,1	0,40	suitable
21	0,99	-0,1	0,40	suitable
34	0,92	-1,0	0,44	suitable
30	0,96	-0,5	0,43	suitable
13	0,86	-1,0	0,45	suitable
23	0,92	-0,5	0,47	suitable
3	0,83	-0,7	0,41	suitable
26	0,90	-0,8	0,48	suitable
39	0,83	-2,1	0,51	suitable
10	0,81	-1,2	0,48	suitable
4	0,79	-1,3	0,49	suitable
11	0,78	-1,7	0,52	suitable
31	0,63	-1,9	0,49	suitable
19	0,84	-2,1	0,53	suitable
18	0,65	-2,4	0,56	suitable
32	0,63	-2,5	0,57	suitable
20	0,75	-3,4	0,61	suitable
8	0,66	-3,4	0,63	suitable
22	0,66	-3,0	0,62	suitable
7	0,61	-3,4	0,63	suitable

Based on the Table 4, there are 6 items (15%) that are not fit, namely items 40, 9, 28, 27, 36, and 37, and the analysis results show that 34 items (85%) are fit. Item fit analysis is used to determine whether the item has functioned normally or not in a measurement. The analysis shows that the item fits the model, so it can be concluded as a valid item. The item fits the model when at least two suitable item criteria are accepted (Sumtono & Widhiarso, 2015). From the response patterns in the table, it can be seen further by looking at the schalogram in Table 5.

Table 5. Student response patterns based on the Scalogram or Guttman matrix

Item																																								Code Person
	3	1	3		1	1	1	2	1		2			3			1	2	3		1	1	2	3	1	2	3	2	3	3	2	3	2	2		2	1	3	4	
3	1	0	2	4	8	4	3	3	1	7	2	2	8	3	1	6	6	1	9	5	7	9	9	7	5	0	6	7	4	0	8	8	5	6	9	4	2	5	0	
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1	1	1	1	0	0	164P
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1	1	1	1	0	0	167P
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1	1	1	1	0	0	169P
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	0	0	0	0	0	074P
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	0	0	0	0	0	075P
1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	120L
1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	122P
1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	123P
1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	125P
1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	127P
1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	0	1	0	0	1	1	0	1	1	1	0	1	0	1	1	0	0	0	0	0	0	147L
1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	0	1	0	0	1	1	0	1	1	1	0	1	0	1	1	0	0	0	0	0	0	151L

Table 5 is a description of the Table 4. The scalogram or Guttman matrix in the data above, from 212 students there are 12 students who have the same answer response pattern. This Scalogram data is arranged based on the easy level to the most difficult level (from item 3 to item 40), as well as simultaneously the respondents are sorted from the lowest rank to the most capable (from person 151 to 164). In table 5, it can be seen further the direct cause if there is an inappropriate response pattern. This table can determine the consistency of students' thinking and can find out if there is fraud committed by students. According to Sumintono & Widhiarso (2015), the scalogram can indicate fraud and guesses the answer in the sample response pattern.

In the Table 5, students with codes 164P, 167P, and 169P have the exact same response which means that these students are cheating on each other. In Table 5 it can be seen that students with code 074P and 075P have the same person logit value (1.48) and have the same response pattern, so that the code 074P and 075P are suspected of cheating each other while doing the test. The total number of students who had a mutual cheating response pattern was 12 students. This shows that some of the students in working on the questions were not in accordance with their respective abilities because they indicated cheating.

Apart from identifying indications of cheating, the Guttman matrix can also indicate guesses. For example in Table 5, the sample code 147L and 151L answered incorrectly for easy questions (item 2) and answered correctly for difficult questions (item 25). The wrong answer given shows that the sample in working on the problem is not careless. Thus, the Guttman Matrix shows that the principle of ordering based on ability and difficulty level of questions is very useful for explaining abilities, even making predictions about a person's ability.

Distribution of Student Ability Level and Difficulty Level of the Item Test. Information about the distribution of the student's ability level and the difficulty level of the items test can be seen on the item person map (wright map). The results of the student's ability level analysis can also be used to see the same abilities of students, namely if the logit value obtained is the same. The results of the analysis can be seen in Figure 3.

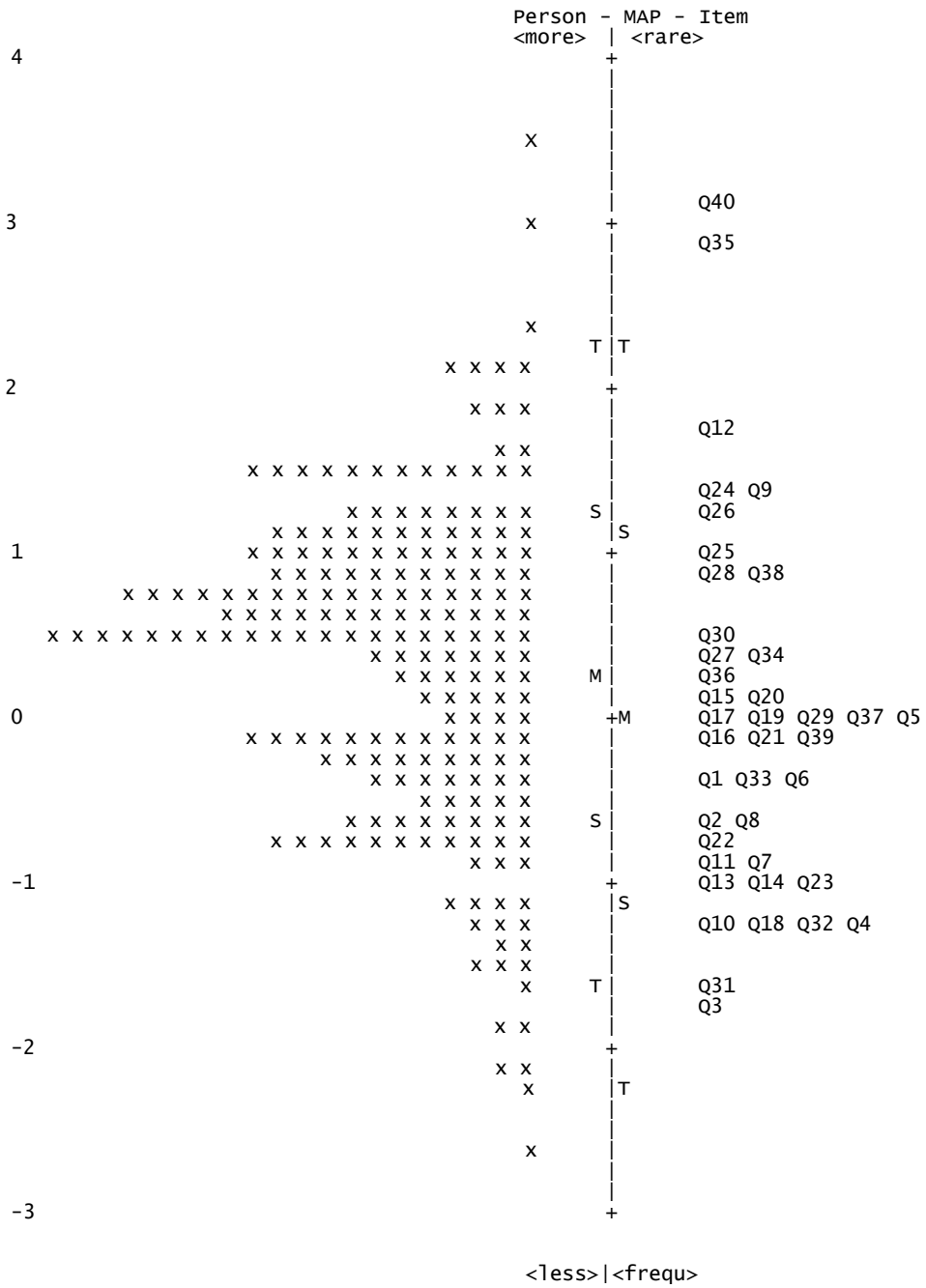

Figure 3. Distribution of respondent ability and test difficulty (person-item map)

Based on Figure 3, it shows that the level of student's ability in answering questions is not much different; this is shown on the map of its position which is close to each other between students. The letter M in Figure 3 shows the mean of person and item (Jüttner, et al., 2013). The map on the left of the diagram shows the distribution of the ability level of student's (respondents) with a person measure of 0.30 logit, while the right side of the diagram shows the distribution of the level of item difficulty, with an item measure of 0.00 logit. The distribution of students' ability levels has a logit price range between the lowest being below -2 (between -2 and -3) and the highest above +3 (between +3 and +4). There

are 139 students (65.6%) who have the ability with a logit price greater than or equal to "0" which indicates that the category is sufficient. Meanwhile, 73 other students (34.4%) had lower abilities than the average difficulty level of the questions (logit value below 0). The logit person value can be seen on the lefthand side, for person 202P with +3.56 logit indicates the person with the highest ability (able to work on almost all questions); person 27P with -2.57 shows students with the lowest ability (at least in solving the questions correctly). While the difficulty level of the questions is seen on the right side of the diagram, the getting to the top it means to be the most difficult question (for example: Q40 questions); while getting down means the easiest problem (example: question Q3).

Various information provided by Wright Map can help educators in evaluating students and item questions. Educators can identify the abilities of individual learners and can also analyze the quality of the questions being tested. In addition, the logit scale has the same interval on the Wright Map, so the information obtained is the right information, for example, educators can find out the number of items that students are unable to do correctly, so that they can make efforts to improve the question items.

The purpose of the ability / difficulty index analysis is to determine the chances of the correct answer to a problem at a certain ability level. The item difficulty parameter is expressed in logit units. Good items have an item difficulty range between -2.0 until +2.0 logit (Hambleton & Swaminathan, 1985). An item is considered as too difficult an item if it has index difficulty above +2.00 logit, whereas if an item has an index difficulty below -2.0 logit it is considered too easy an item. This study refers to the interpretation of the item difficulty value used by Adedoyin & Mokobi (2013) which is categorized as very difficult if the value of b (measuring item)> 1; difficult $0.5 \leq b < 1$; moderate $-0.5 \leq b < 0.5$; easy $-0.5 \leq b < -1$; and very easy $b \leq -1$. The results of the item difficulty level on the instruments can be seen in Table 6.

Table 6. Item difficulty level

Item number	Score	Difficulty Index (Measure)	Category
40*	17	3,09	Very difficult
35	20	2,90	Very difficult
12	49	1,72	Very difficult
9*	61	1,38	Very difficult
24	61	1,38	Very difficult
26	64	1,30	Very difficult
25	77	0,98	Difficult
38	79	0,93	Difficult
28*	80	0,91	Difficult
30	97	0,51	Difficult
34	103	0,38	Moderate
27*	105	0,33	Moderate
36*	109	0,24	Moderate
20	112	0,18	Moderate
15	116	0,08	Moderate
19	117	0,06	Moderate
29	117	0,06	Moderate
37*	117	0,06	Moderate
5	120	-0,01	Moderate
17	120	-0,01	Moderate
21	123	-0,08	Moderate
39	123	-0,08	Moderate
16	124	-0,10	Moderate
6	134	-0,34	Moderate
1	136	-0,39	Moderate
33	138	-0,43	Moderate
2	145	-0,61	Easy
8	145	-0,61	Easy
22	152	-0,80	Easy
7	155	-0,88	Easy
11	156	-0,91	Easy
23	157	-0,94	Easy
13	158	-0,97	Easy
14	159	-1,00	Very easy
4	166	-1,22	Very easy
18	166	-1,22	Very easy
10	167	-1,25	Very easy
32	167	-1,25	Very easy
31	178	-1,65	Very easy
3	181	-1,77	Very easy
Mean	**119,3**	**0,00**	
S.D.	**41,2**	**1,09**	

Note: *) indicates the item is not fit

Based on the Table 6, the results of the item difficulty index are well distributed in the categories very easy (7 items = 17.5%), easy (7 items = 17.5%), moderate (16 items = 40%), difficult (4 items = 10%), and very difficult (6 items = 15%) with a difficulty index range from 3.09 to -1.77. Based on this range value, it can be said that the instrument item has a good difficulty index.

In Table 6, you can see several columns that provide information about each item. This table is sorted according to the level of difficulty, which is based on the measure value which is the logit value of each item. The highest logit value indicates a difficult question. This corresponds to the total score column, which states the number of correct answers. For example, for the 40th question the logit score was 3.09 logit and only 17 students answered correctly and it was categorized as difficult questions. Whereas for the 3rd question it has a logit value of -1.77 logit and as many as 181 students who answered correctly, and are included in the very easy question category. The level of ability of students can be seen in Table 6 by looking at the measure value, the higher the measure value, the higher the level of ability of students and conversely the lower the measure, the lower the level of ability of students.

Identification of Item Bias with DIF (Differential Item Functioning) Test. Item bias is a test condition that is unfair, inconsistent, and polluted by factors outside the ability factor to be measured. Item bias result in a test that is discriminatory or in favor of certain groups whose causes can be viewed from various aspects that have absolutely nothing to do with ability factors, such as gender, ethnicity, culture, region, and others (Osterlind, 1983; Chan & Subramaniam, 2020). So that the bias of a test can be interpreted as invalidity or systematic error in measuring members of a group under study. The plot results of items identified by DIF or indications of item bias can be seen in Figure 4.

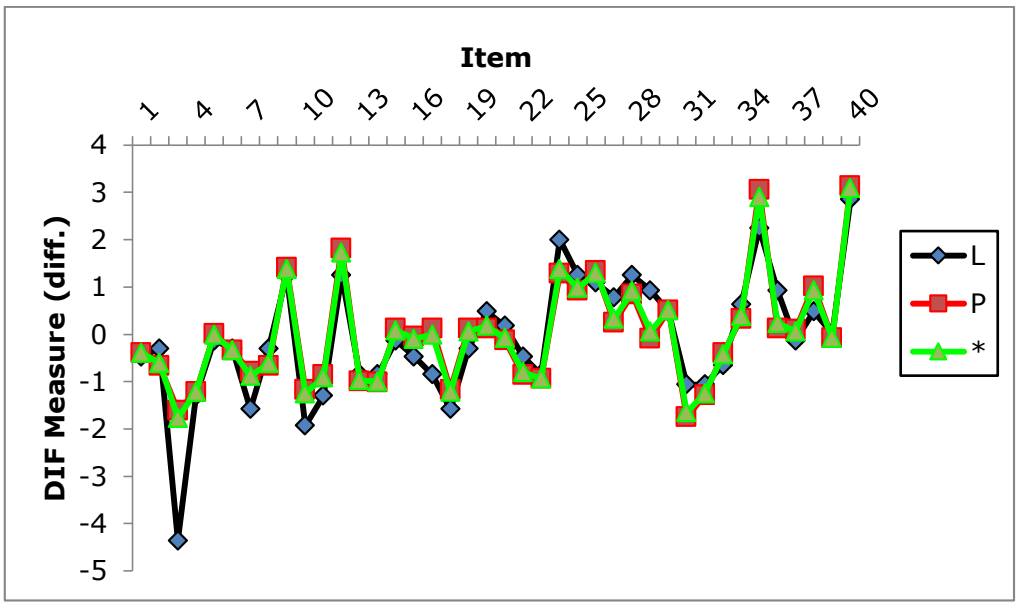

Figure 4. Item bias graph with DIF test

The graph in the Figure 4 shows the relative item difficulty level for each group. The higher the point graph, the more difficult the item is for the group (Osterlind, 1983). There are three curves based on gender, namely L (male), P (female), and an * (star) which shows the average value. From the graph, it can be seen roughly that the distance between the DIF measure values between L and P is the farthest in item number 3. While for other items the distance between L and P is not too far. This shows that in item 3 the difference in the level of difficulty between men and women is quite large. In this case, men benefit more because the item seems more difficult for women than men. Therefore, item number 3 should be reviewed whether it is true that this item is more beneficial for men than women.

Instrument Reliability. Further determination of instrument reliability by paying attention to person separation and item separation. Person separation is an estimate of the extent to which participants can distinguish person on the measured variable, while item separation estimates the extent to which participants can answer all difficulty levels of the item. A reliability index that is higher than 2 is declared satisfactory. Lower reliability scores indicate redundancy in the item and less person variability on the trait. The results of the person and item reliability are presented in Table 7.

Table 7. Summary of instrument statistics: person and item reliability

Criteria	Parameter (N)	
	Persons (212)	*Items* (40)
MNSQ	1,05	1,05
ZSTD	0,0	0,1
Separation	2,23	6,21
Reliability	0,83 (Good)	0,97 (Special)
Alpha Cronbach	0,85 (Good)	

Based on the results of the analysis presented in Table 7, it can be seen that the reliability of the person is in the good category and the reliability of the items is in the special category. The overall interaction between person and item is seen from the alpha Cronbach value (Hayati & Lailatussaadah, 2016; Sumintono & Widhiarso, 2015) with a value of 0.85 which has good criteria. A reliability value greater than 0.80 is a value that indicates high reliability and there is a reliable interaction between person and item (Bond & Fox, 2007; Linacre, 2016; Setiawan, et al., 2018). In other words, the results show the suitability between the person and item relationships used. This item and respondent reliability test shows that this research instrument can be used to measure the dimensions of the construction of students' ability assessments (Yasin, et al., 2018).

The separation value of person (2.23) and item (6.21) is also good category, because according to Linacre (2016), the separation index value is said to be good if it is greater than 2.0 and the separation person and item index is an additional measurement very important for the evaluation of a measuring instrument (Boone, et al., 2014). Separation reliability (item or person reliability) is categorized as high value because the research sample and the difficulty level of the item have a wide range and results in small measurement errors (Lia, et al., 2020). This indicates that the item has a difficulty level from easiest to most difficult. In the research sample, a broad sample means that the sample can be spread from the smartest to the least intelligent (Linacre, 2016).

The information function describes how well each ability level can be estimated (Baker & Kim, 2017). The information function is used for interpreted as reliability in classical test theory, but it is more accurate to estimate the latent nature of the responden than the reliability coefficient (Rosana, et al., 2020). The maximum IF value of the initial chemical ability of an instrument with 40 items was found to be 8.21429. The information function graph indicates that the statements used are not too difficult (small logit values) and can provide good information for participants with slightly lower abilities than participants with moderate abilities.

Conclusion

Based on the discussion that has been described, it can be concluded that: The quality of the Chemistry National Examination (UN) questions is categorized as very good based on the following aspects: (1) Unidimension is 28.6% and the variance-covariance matrix is close to 0; (2) 36 items indexed as fit item (85%); (3) Item difficulty level in order from easy: medium: difficult (35%:40%:25%); (4) Person Measure 0.30 logit shows the average ability of the respondents above the average item; (5) The score of logit person 202P (+3.56 logit) shows the person with the highest ability; person 27P (-2.57 logit) indicates the respondent with the lowest ability; (6) item number 3 there is a gender bias, this item appears to be more difficult for women than for men; (7) Person Reliability is 0.83 while Item Reliability is 0.97 with a Cronbach α of 0.85 which is a good criteria.

References

Adedoyin, O. & Mokobi, T. 2013. Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. *International Journal of Asian Social Science*, 3(4):992-1011.

Afrassa, T.M. 2005. *Monitoring mathematics achievement over time. In Applied Rasch Measurement: A Book of Exemplars*, Springer, Dordrecht, pp.61-77.

Ardiyanti, D. 2016. Aplikasi model rasch pada pengembangan skala efikasi diri dalam pengambilan keputusan karir peserta didik. *Jurnal Psikologi*, 43(3):248–263.

Baker, F.B. & Kim, S.H. 2017. *The Basics of Item Response Theory using R*. Springer, New York, pp.55-67.

Bond, T.G. & Fox, C.M. 2007. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences,* 2nd Edition, Lawrence Erlbaum Associates, Publisers, Mahwah, New Jersey, London.

Boone, W.J., Staver, J.R., & Yale, M.S. 2014. *Rasch Analysis in the Human Sciences*, Springer Science dan Business Media.

Brown, R.L., Obasi, C.N., & Barrett, B. 2016. Rasch analysis of the WURSS-21 dimensional validation and assessment of invariance. *Journal of Lung, Pulmonary & Respiratory Research*, 3(2):46-53.

Chan, S.W., Ismail, Z., & Sumintono, B. 2014. A rasch model analysis on secondary students' statistical reasoning ability in descriptive statistics. *Procedia-Social and Behavioral Sciences*, 129:133-139.

Chernyshenko, O.S., Stark, S., Chan, K.Y., Drasgow, F., & Williams, B. 2001. Fitting item response theory models to two personality inventories: issues and insights. *Multivariate Behavioral Research,* 36(4):523-562.

Darmana, A., Jasmidi, & Sutiani, A. 2020. Development of the thermochemistry-Hots-tawheed multiple choice instrument. *In Journal of Physics Conference Series*, 1462(1):1-9.

Field, A. 2009. *Discovering Statistics using SPSS*, 3rd edition, Sage Publication Ltd, London.

Greiff, S., Fischer, A., Wüstenberg, S., Sonnleitner, P., Brunner, M., & Martin, R. 2013. A multitrait–multimethod study of assessment instruments for complex problem solving. *Intelligence*, 41(5):579-596.

Guler, N., Uyanik, G.K., & Teker, G.T. 2014. Comparison of classical test theory and item response theory in terms of item parameters. *European Journal of Research on Education*, 2(1):1-6.

Hambleton R.K. & Swaminathan, H. 1985. *Items Response Theory: Principles And Application,* Kluwer-Nijjhoff Publish, Boston.

Hayati, S. & Lailatussaadah, L. 2016. Validitas dan reliabilitas instrumen pengetahuan pembelajaran aktif, kreatif dan menyenangkan (PAKEM) menggunakan model Rasch. *Jurnal Ilmiah Didaktika: Media Ilmiah Pendidikan dan Pengajaran*, 16(2):169-179.

Ibrahim, F.M., Shariff, A.A., & Tahir, R.M. 2015. Using rasch model to analyze the ability of pre-university students in vector. *In AIP Conference Proceedings*, 1682(1): 030009.

Isnani, I., Utami, W.B., Susongko, P., & Lestiani, H.T. 2019. Estimation of college students' ability on real analysis course using rasch model. *REiD (Research and Evaluation in Education)*, 5(2):95-102.

Jüttner, M., Boone, W., Park, S., & Neuhaus, B.J. 2013. Development and use of a test instrument to measure biology teachers' content knowledge (CK) and pedagogical content knowledge (PCK). *Educational Assessment, Evaluation and Accountability*, 25(1):45-67.

Kaiser, H.F. 1974. An index of factorial simplicity. *Phyhcometrika*, 39(1):31-36.

Lia, R.M., Rusilowati, A., & Isnaeni, W. 2020. NGSS-Oriented chemistry test instruments: validity and reliability analysis with the rasch model. *REiD (Research and Evaluation in Education)*, 6(1):41-50.

Linacre, J.M. 2016. *A User's Guide to WINSTEPS MINISTEP Rasch-Model Computer Programs*, IL: Winsteps.com, Chicago.

Mahmud, Z., Ramli, W.S.W., Sapri, S., & Ahmad, S. 2017. Diagnosis of students' ability in a statistical course based on rasch probabilistic outcome. *In AIP Conference Proceedings*, 1836(1):020048.

Osterlind, S.J. 1983. *Test Item Bias*, CA: Sage Publication Inc, Beverly Hills.

Pratama, D. & Husnayaini, I. 2020. Applying rasch model to measure students reading comprehension. *JISAE: Journal of Indonesian Student Assessment and Evaluation*, 6(2):203-209.

Reckase, M.D. 1979. Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics*, 4(3):207-230.

Rosana, D., Widodo, E., Setianingsih, W., & Setyawarno, D. 2020. Developing assessment instruments of PISA model to measure students' problem-solving skills and scientific literacy in junior high schools. *Jurnal Pendidikan Sains Indonesia*, 8(2):292-305.

Sabekti, A.W. & Khoirunnisa, F. 2018. Penggunaan rasch model untuk mengembangkan instrumen pengukuran kemampuan berpikir kritis peserta didik pada topik ikatan kimia. *Jurnal Zarah*, 6(2):68–75.

Saidi, S.S. & Siew, N.M. 2019. Reliability and validity analysis of statistical reasoning test survey instrument using the rasch measurement model. *International Electronic Journal of Mathematics Education*, 14(3):535-546.

Samritin, S., Wijaya, R.S., Tarno, T., Suranata, K., Ardi, Z., Ifdil, I., ... & Rangka, I.B. 2019. Matching the student's ability and their math test using rasch analysis. *In Journal of Physics: Conference Series*, 1318(1):012059.

Setiawan, B., Panduwangi, M., & Sumintono, B. 2018. A Rasch analysis of the community's preference for different attributes of islamic banks in Indonesia. *International Journal of Social Economics*, 45(12):1647-1662.

Sihombing, R.U., Naga, D.S., & Rahayu, W. 2019. A Rasch model measurement analysis on science literacy test of indonesian students: smart way to improve the learning assessment. *IJER-Indonesian Journal of Educational Review*, 6(1):44-55.

Smits, N., Cuijpers, P., & Van Straten, A. 2011. Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research*, 188(1):147-155.

Sumintono, B. & Widhiarso, W. 2015. *Aplikasi Pemodelan Rasch pada Assessmen Pendidikan,* Trim Komunikata, Cimahi.

Susongko, P. 2016. Validation of science achievement test with the rasch model. *Jurnal Pendidikan IPA Indonesia*, 5(2):268-277.

Wati, M., Mahtari, S., Hartini, S., & Amelia, H. 2019. A Rasch model analysis on Junior High School students' scientific reasoning ability. *International Journal of Interactive Mobile Technologies (iJIM)*, 13(7):141-149.

Wibisono, S. 2019. Aplikasi model rasch untuk validasi instrumen pengukuran fundamentalisme agama bagi responden muslim. *JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)*, 3(3):729-750.

Wu, Q., Zhang, Z., Song, Y., Zhang, Y., Zhang, Y., Zhang, F., & Miao, D. 2013. The development of mathematical test based on item response theory. *International Journal of Advancements in Computing Technology,* 5(10):209–216.

Yasin, S.N.T.M., Yunus, M.F.M., & Ismail, I. 2018. The Use of rasch measurement model for the validity and reliability. *Journal of Counseling and Educational Technology*, 1(2):22-27.