Tersedia online di EDUSAINS
Website: http://journal.uinjkt.ac.id/index.php/edusains

EDUSAINS,14(1), 2022, 24-33

Research Artikel

VALIDATION OF ASSESSMENT INSTRUMENTS FOR INTEGRATED SCIENCE LEARNING ON THE ABILITY OF STUDENT USING RASCH MODEL

Marantika Lia Kristyasari[1*], Sri Yamtinah[2]
[1] Program Studi Pendidikan Kimia, FKIP Universitas Musamus, Indonesia
[2] Program Studi Pendidikan Kimia, FKIP Universitas Sebelas Maret, Indonesia
marantika_fkip@unmus.ac.id

Abstract

The Computerized Two Tier Multiple Choice (CTTMC) instrument has been created to mesure the bility of student in Integrated Science learning. This research was conducted to produce empirical evidence regarding the validity of the TTMC instrument using the Rasch Model analysis thorugh the Winstep program. A total of 283 junior high school students in the city of Surakarta were the research subjects. This assessment instrument consists of 20 questions. Statistical analysis was performed using the Rasch model. The results of the analysis show that in general, this assessment instrument can explain 40.7% of the variance that appears in each group of respondents. As many as 85% of the questions were declared fit and 15% of the questions still needed to be improved. Of the 283 participants, 13,4% had an ability misfit, 13,8% were included in the upper outlier, and 72,8% of the participants had an infit ability. Thus, that the assessment of Integrated Science SMP learning, in general, can be carried out using this instrument.

Keywords: *Assessment instrument; integrated science learning; RASCH model; student ability; validation.*

Permalink/**DOI:** http://doi.org/10.15408/es.v13i2.22468

How To Cite: Kristyasari, M.L., Yamtinah, S. (2022). Validation Of Assessment Instruments For Integrated Science Learning On The Ability Of Student Using Rasch Model. *EDUSAINS*, 14 (1) : 24-33.

*Corresponding author

Received: 05 May 2021; Revised: 12 January 2022; Accepted: 15 July 2022

INTRODUCTION

Evaluation tools also have an important role in the learning process as well as learning methods and models. The results of the assessment can be used as a benchmark for teachers and students to determine the success of the learning process that takes place in the classroom. There are two types of evaluation tools used by teachers, namely in the form of written and oral tests. An evaluation tool in the form of an oral, is usually used by teachers to measure the readiness of students in receiving new material and to measure students for the material that has been given at the previous meeting. Written test is a test that is often used by teachers as a tool to measure student learning outcomes (Mardapi, 2016).

The CTTMC instrument is a combination of multiple choice questions and description questions, so that the purpose of using this instrument is to maximize student learning outcomes, reveal the knowledge possessed by students in more depth, and can be used practically without fear of subjectivity in assessment and can also be used in practice. reduce acts of cheating committed by students at the time of answering (lucky guest/guesing). This instrument was developed by the author using a computer system of 20 questions accompanied by individual profiles of students. This individual student profile contains reports on the results of student answers, the value of each item, the value of student learning outcomes, both knowledge and abilities of other students.

A good and appropriate assessment instrument to be used is an instrument that can provide accurate information related to the ability of students on the competencies being tested. The focus of this research is to test the feasibility (validity) of the integrated science learning assessment instrument given to class VIII junior high school students whose data is then analyzed using the Rasch model.

The Rasch model is a modern valuation theory that can classify item and person calculations in a distribution map (Rozeha, Azami, 2007). The Rasch model is based on two principles, namely the ability of students and the relationship between students' abilities and the level of difficulty of the items (Aprilia et al., 2021; Sumintono & Widhiarso, 2015; Wigati & Kurratul, 2020). Rasch can analyze the quality of items such as validity and is very effective in identifying students' conceptual abilities in detail (Bohori & Liliawati, 2019; Rusmansyah & Almubarak, 2020). The Rasch measurement model is formed from the ability of each respondent (student) who answers the test with the difficulty of each test item being tested (Yasin et al., 2018). His research (Winarti & Mubarak, 2019)also states that the Rash model can provide a concrete and comprehensive description of measuring assessment instruments. This is because the Rasch model involves two parameter aspects, namely the ability of students and the level of difficulty of the questions (Mariam et al., 2018).

Therefore, the purpose of this study was to determine the feasibility of the CTTMC assessment instrument by looking at the relationship between the abilities of students and the level of difficulty of the items from the results of the analysis using the Rasch model.

METHODS

This assessment uses descriptive quantitative methods that aim to obtain information about the feasibility of the items using the Rasch model analysis. The Rasch model emphasizes that every student has the same opportunity to answer questions correctly. Questions that have different levels of difficulty in Rasch are called person logit and item logit (Misbah et al., 2019; Sihombing et al., 2019).

The subjects in this study were students of class VII SMP Negeri 1, 4, and 12 in the city of Surakarta, amounting to 283 people. Data collection was carried out in this study using the test method. Assessment instrument developed by the author on the Integrated

Natural Sciences material. The preparation of the question grid begins at the stage of making a concept map to see the integration of science from the three existing science aspects, analyzing the integrated science syllabus found in class VII SMP, dissecting the contents of basic competencies which are then revealed to be indicators of basic competence and the last is to make a question indicator as many as 20 items. The data obtained from this study are in the form of response patterns of students' answers based on the results of doing tests using the CTTMC instrument. The data was then analyzed using the Rasch model analysis.

The analysis of the Rasch model can provide information on the existence of individuals or respondents who have inappropriate response patterns and invalid questions or are often referred to as outliers/misfits. There are three criteria used in checking the suitability of items that are in the outliers/misfit category and respondents who have an inappropriate response pattern (not fit) (Boone et al., 2014) namely the MNSQ value: 0.5 < MNSQ < 1.5; ZSTD Outfits: -2.0 < ZSTD < +2.0; and Pt Measure Corr: 0.4 < Pt Measure Corr < 0.85. If there are items that do not meet the three criteria, then the items are declared to have poor quality so that improvements (revisions) need to be made. This is because the ability of students who are tested must go through good quality items. The difficulty level of the items can be categorized based on the average logit value and the standard deviation value on the item measure. The category of group questions based on the level of difficulty of the items can be seen in Table 1.

Table 1. Categories of question groups based on the level of difficulty of the items

Logit value	Category
Greater than +1,09 SD	***Very difficult***
0,0 logit +1,09 SD	***Difficult***
0,0 logit –1,09 SD	***Currently***
Smaller than –1,09 SD	***Easy***

(Sumintono & Widhiarso, 2015)

The person measure category uses the SD Standard Deviation value). The criteria for grouping the abilities of students are listed in Table 2.

Table 2. Criteria for Grouping Student Abilities

Logitability value of students	Category
Greater than +1,11	High
Smaller than +1,11	Medium
Smaller than –0,28	Low

(Sumintono & Widhiarso, 2015)

RESULT AND DISCUSSION

Data analysis was carried out using the ministep winsteps 3.73 software model. The ability to process data using Winsteps is higher than the usual MiniStep software. Ministep software can only analyze 75 students and 25 questions, but Winsteps software can analyze hundreds or even thousands of subjects (Sumintono & Widhiarso, 2015). In this study, the CTTMC assessment instrument developed was 20 items with 283 respondents.

Summary of Statistics

Summary of Statistics Instruments for both item and person measurements are presented in Table 3.

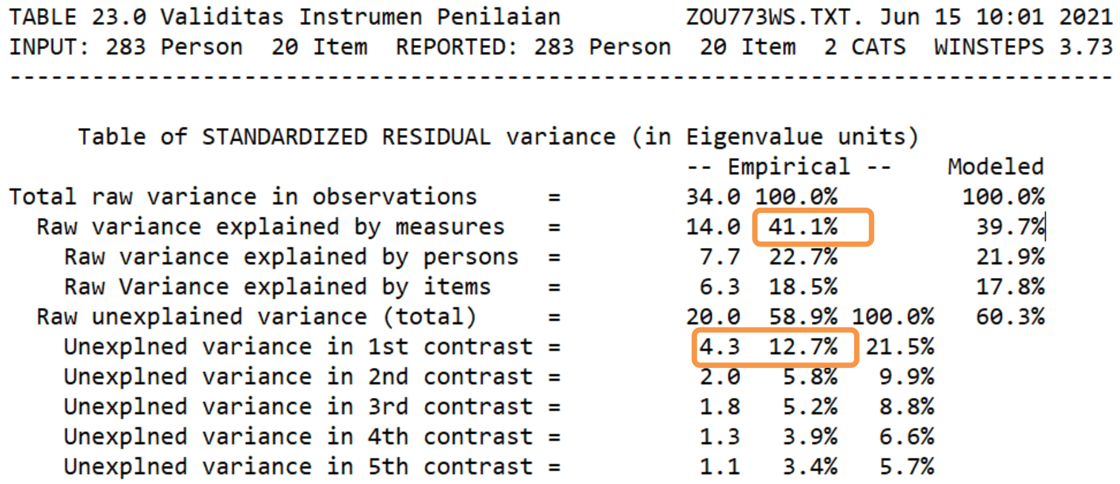
Table 3. Summary of instrument statistics in terms of students and questions

	Mean logit value
Students/responden	-0,18
Question items	0,00

Table 3. shows the results that the mean logit value of the students (respondents) is -0,18 while the mean logit value for the questions is 0.00. This shows that each item has no group differences, while the respondents have group differences and even many groups are outliers.

Validity

Instrument validity to examine instrument ability in Integrated Science evaluation toward student ability. In evaluation, validity always emphasizing the depth of the item's ability and suitability against the student's ability using a specific concept or settle concept definition (Djaali, H., Pudji M., Sudarmanto, 2008). In Rasch analysis, the validity test is often known as Item Undimensionality (Sumintono & Widhiarso, 2015). Item Undimensionality is the ability measurement of evaluation instrument which developed and proclaimed valid for evaluating learning activity. Rasch analysis using analysis which has principal component analysis from standardized residual variance (in Eigenvalue units) (Sumintono & Widhiarso, 2015). Validity test based on Item Undimensionality shown in Table 3. Output result showing usability test instrument items e.g. which part suitability of the items instrument for measuring student's ability. Instrument validity analysis is declared as a fit test or misfit. Item Undimensionality result can be seen in Picture 1.



```
TABLE 23.0 Validitas Instrumen Penilaian          ZOU773WS.TXT. Jun 15 10:01 2021
INPUT: 283 Person  20 Item  REPORTED: 283 Person  20 Item  2 CATS  WINSTEPS 3.73
-------------------------------------------------------------------------------

     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                                -- Empirical --    Modeled
Total raw variance in observations     =        34.0 100.0%         100.0%
  Raw variance explained by measures   =        14.0  41.1%          39.7%
    Raw variance explained by persons  =         7.7  22.7%          21.9%
    Raw Variance explained by items    =         6.3  18.5%          17.8%
  Raw unexplained variance (total)     =        20.0  58.9% 100.0%   60.3%
    Unexplned variance in 1st contrast =         4.3  12.7%  21.5%
    Unexplned variance in 2nd contrast =         2.0   5.8%   9.9%
    Unexplned variance in 3rd contrast =         1.8   5.2%   8.8%
    Unexplned variance in 4th contrast =         1.3   3.9%   6.6%
    Unexplned variance in 5th contrast =         1.1   3.4%   5.7%
```

Picture 1. Output *Item Undimensionality* in Winstep

In Picture 1, output result showed at raw variance explained by measures column in empirical part. Itemundimensionality criteri a interpretation is divide into 3 part, ie. 1) score > 20% including to fulfilled validation criteria; 2) > 40% is good; and 3) > 60% including outstanding validation criteria. Other can be seen in eigenvalue and observed value whether some items are problematic or unsuitable. With criteria in unexplained variance, 1st contrast and eigenvalue must less than 3. It shows whether there are problematic items for the observed value must less than 15% to show the suitable item (item fit). The Winstep software’s validity processing result is presented in Table 4.

Table 4. Instrument Validity Processing Result

Raw variance explained by measures	Interpretation	Unexplained variance 1st contrast		Interpretation
		Eigenvalue	Observed	
41,1%	Good	4,3	12,7%	There is a problem item

According to Table 4. The results obtained from *raw variance explained by measures* value showing that overall evaluation instrument item is in “good” category. Other for *an observed* score in *unexplained variance 1st contrast* result is shows that there is no tendency *to misfit* items. Because the obtained percentage is 12,7% less than 15%, so 20 items are used. But for *the eigenvalue* score gain 4,3 more than 3, this indicates that there are problem items. So, need further analysis using *item fit order* analysis which aims to determine that an item needs to be repaired, replaced, or still maintained.

Item fit is an item's suitability that explains if the item normally operates or not when taking measurements. There are three criteria for determining the suitability (*fit* or *misfit*)of an item (Sumintono & Widhiarso, 2015), that is:

1. *Outfit Mean Square* (MNSQ) accepted value is: 0,5 < MNSQ < +1,5

2. *Outfit Z-Standard* (ZSTD) accepted value is: -2,0 < ZSTD < +2,0
3. *Point Measure Correlation* (Pt Mean Corr) accepted value is: 0,4 < Pt Mean Corr < +0,85

Processing result of suitability test and whether an item is suitable or not can be see in Table.5.

Tabel 5. Result Fit/Misfit Items Test

Entry Number	Total Score	Total Count	Measure	Mode S.E	INFIT		OUTFIT		PT-MEASURE		EXACT MATCH		Item
					MNSQ	**ZSTD**	**MNSQ**	**ZSTD**	**CORR.**	**EXP.**	**OBS%**	**EXP %**	
1	72	283	2,47	0,21	1,60	4,2	3,92	4,7	A. 0,52	0,72	83,2	87,9	S1
12	181	283	-0,72	0,16	1,94	9,9	2,21	6,1	B. 0,28	0,58	44,3	75,7	S12
11	174	283	-0,54	0,16	168	7,4	1,83	4,8	C. 0,39	0,59	52,0	76,2	S11
9	217	283	-1,66	0,17	1,34	4,1	1,81	2,9	D. 0,35	0,48	78,3	77,3	S9
13	170	283	-0,44	0,16	0,97	-0,4	1,48	3,1	E. 0,59	0,60	79,9	76,5	S13
17	124	283	0,75	0,17	1,39	4,1	1,33	1,8	F. 0,57	0,67	66,8	78,9	S17
15	107	283	1,23	0,17	1,34	3,6	1,34	1,5	G. 0,60	0,69	70,9	79,7	S15
20	129	283	0,61	0,16	1,22	2,5	1,30	1,8	H. 0,60	0,67	74,6	78,7	S20
6	123	283	0,77	0,17	1,23	2,5	1,25	1,4	I. 0,61	0,68	70,9	79,0	S6
10	251	283	-2,83	0,21	0,99	0,0	1,21	0,6	J. 0,33	0,36	91,0	88,2	S10
3	131	283	0,56	0,16	0,80	-2,5	0,63	-2,7	j. 0,73	0,67	81,1	78,6	S3
14	147	283	0,14	0,16	0,73	-3,5	0,65	-2,9	i. 0,72	0,64	82,4	77,5	S14
2	178	283	-0,64	0,16	0,70	-4,5	0,56	-3,6	h. 0,68	0,59	85,7	76,0	S2
7	143	283	0,24	0,16	0,66	-4,6	0,54	-4,0	g. 0,75	0,65	87,3	77,7	S7
4	183	283	-0,77	0,16	0,62	-6,0	0,48	-4,2	f. 0,70	0,57	87,3	75,7	S4
19	136	283	0,42	0,16	0,58	-5,9	0,44	-4,8	e. 0,79	0,66	86,9	78,2	S19
8	148	283	0,11	0,16	0,57	-6,2	0,50	-4,5	d. 0,77	0,64	91,0	77,5	S8
16	157	283	-0,12	0,16	0,55	-6,6	0,47	-5,0	c. 0,76	0,63	93,0	77,1	S16
18	147	283	0,14	0,16	0,52	-7,1	0,43	-5,4	b. 0,79	0,64	90,6	77,5	S18
5	142	283	0,27	0,16	0,49	-7,5	0,39	-5,7	a. 0,80	0,65	91,8	77,7	S5
MEAN	153,0	283,0	0,00	0,17	0,99	-0,8	1,14	-0,7			79,4	78,6	
S.D.	38,0	0,0	1,06	0,02	0,43	5,2	0,84	3,8			12,8	3,3	

Based on the fit item test above, the result obtained several items have unsuitable criteria of *Outfit* MNSQ, *Outfit* ZSTD, dan Pt Mean Corr value. Three items not fulfilling the criteria at all, ie. Item number S12, S11, and S9. Item number S12 has *Outfit* MNSQ, *Outfit* ZSTD, dan Pt Mean Corr value that is 2,21 (>1,5); 6,1 (>2,0); and 0,28 (<0,4). While item number S11 *Outfit* MNSQ, *Outfit* ZSTD, dan Pt Mean Corr value is 1,83 (>1,5); 4,8 (>2,0); and 0,39 (<0,4). Item number S9 has *Outfit* MNSQ, *Outfit* ZSTD, dan Pt Mean Corr value that is 1,81 (>1,5); 2,9 (>2,0); and 0,35 (<0,4). *Outfit* MNSQ value of the three items is > 1,5 it identifies the three items data as a complex item that needed further understanding. So, it raises the diversity of student answers variation who cause the three items are not suitable or deviate from the model used as an analytical reference. Therefore, so that three items are compatible with the model it needs improvement and can fulfill criteria requirements as a fit item.

Besides that, four items have *Outfit* MNSQ score < 0,5 and ZSTD < -0.2, it is item numbers S4, S19, S16, and S18. The score of item number S4 is 0,48 (< 0,5); and -4,2 (<-2,0). In item number S19, score obtained is 0,44 (< 0,5); and -4,8 (<- 2,0). Item number S16's score is 0,47 (< 0,5); and -5,0 (< -2,0) also item number S18's score is 0,43 (<

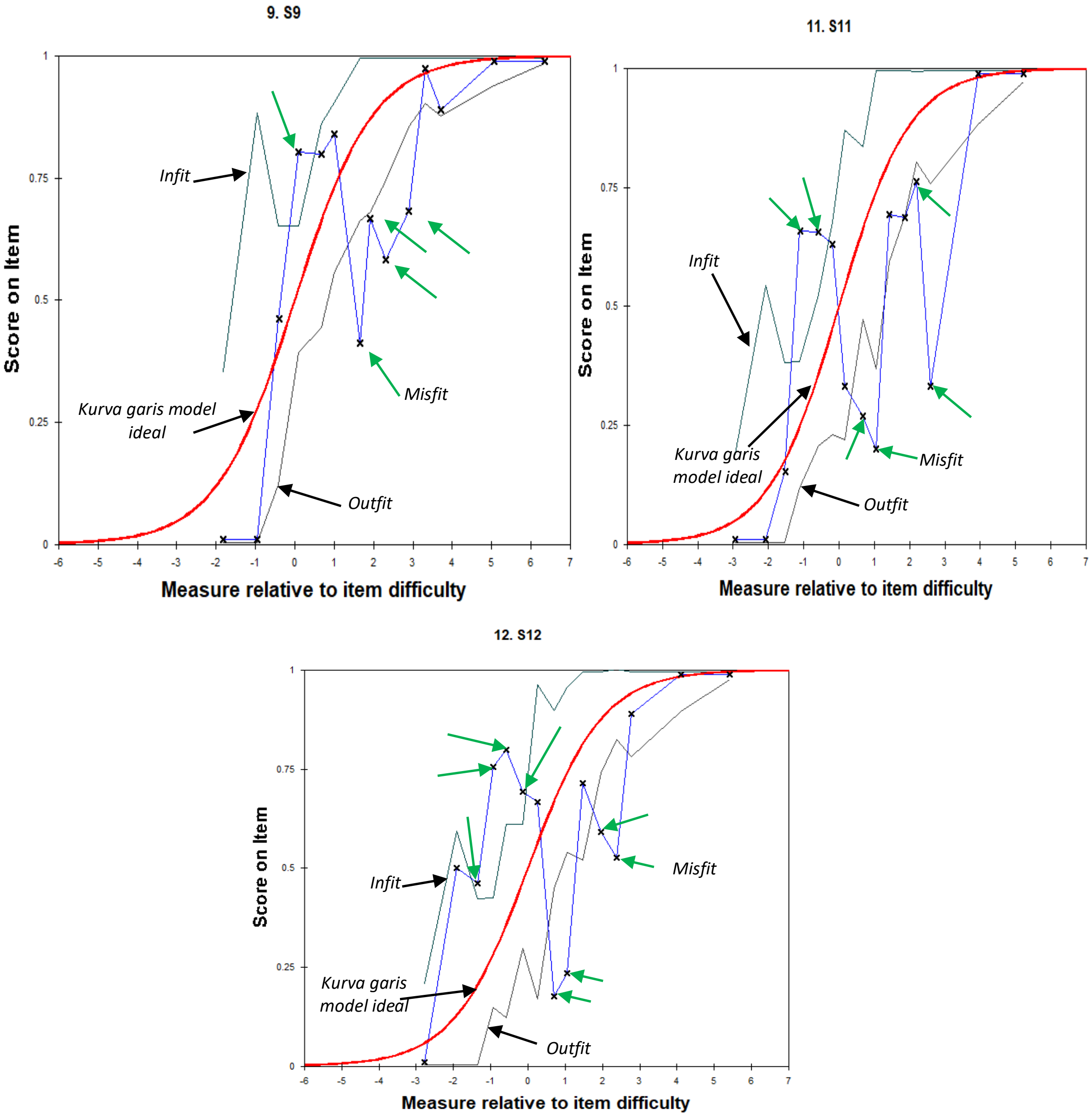
0.5); and -5,4 (< -2,0). *Outfit* MNSQ score < 0,5 indicating that the observed data has less 22% data variation than model Rasch's data prediction (T. Bond, 2015). While *the Outfit* ZSTD score is negative, it indicates that the data has slight variation compared to the Rasch model data used. In other words, the response given by the student is close to *the Guttman-style response string* model. *A Guttman-style response string* is a response created from the subject condition or student who has the high ability which can answer the items correctly, and students who have the low ability who cannot answer the items correctly. However, the four items are including a fit item group and can be used.

An item, its *Outfit* MNSQ, and ZSTD score is more than established criteria, but for Pt Mean Corr score is in score range that suitable for criteria, this item is number S1. Item number S1*Outfit* MNSQ and ZSTD score are 3,92 and 4,7 it is the highest score of all analyzed items. This happens because the obtained data from the items have at the most variation than other items. While for Pt Mean Corr's score is obtained by 0,52. Even if item number S1 is in the MNSQ and ZSTD category has more response variation than the model used, but in Pt Mean Corr score, this item is still stable. In other words, even has more response variation, student response consistency is still maintained because item S1 can distinguish the discriminatory power item well. So, item S1 can be claimed as a fit item and can continue to be used for the evaluation item of Integrated Science Learning.

There are items do not meet the requirements in one of the criteria, such as item number S2, S3, S7, S13, and S14. This items group does not meet the requirement on its *Outfit* ZSTD score is < -2,0. However, because the amount of ZSTD score depends on the number of samples used and does not meet the requirement of only one score, then this item group can be claimed as fir items. It is consistent with the theory proposed by (Sumintono & Widhiarso, 2015) if items do not meet only one criterion, then the item does not need to be changed or replaced. The item is still being used and proclaimed as a fit item.

It concluded that the misfit item test results showed in Table 3. Of 20 items tested, 17 items were proclaimed as qualified (*fit*) to be continued and used as evaluation instruments in Integrated Science Learning with a note that 3 items need improvement. In other words, these three items are misfit items. This corresponding with the results shown in Table.1 regarding the instrument statistical summary that stated the logit value obtained on the items is 0,0. It shows that overall, the instrument can measure what the purpose of measurement is. The average item logit value is 0.0 is a random value set to express a probability (50:50) in an equivalent measuring between the respondent's ability and the difficulties of the item (T. G. Bond & Fox, 2007).

Other than the number shown in Table. 5. Items Misfit test can be performed using *the expected score ICC* graph. This graphic intended to find out the misfit items. The ICC graph is present in the Picture. 2.



Picture 2. *Expected Score* ICC Item S9, S11, and S12 Graphic

Picture 2 presented *the expected score* ICC graphic it shows that items S9, S11, and S12 are proven if it does not meet the criteria requirements as a fit item. It shows some numbers of *misfit* responses resulted in the three items. There are more than 4 *misfit* responses. In S9 item there is 5 *misfit* response which is outside of the *outfit*. While item S11, 6 *misfit* responses found, it spread outside the *infit* area as many as 2 people and outside the *outfit* area as many as 4 people. Item S12 has the most *misfit* responses that are 8 responses spread outside *the infit* area and outside *outfit* area, each one is 4 responses. The number of *misfit* responses causes the item to have an *Outfit* MNSQ, ZSTD, and Pt Mean Corr value outside the limits of the specified criteria. Thus, the more item that gets *misfit* responses, the items mentioned as unfit items and must be changed or replaced.

Meanwhile, suitability and non-suitable measurement using the same criteria as the item. The processing results of the fit/misfit person test are presented in Table.6.

Table 6. Result of Fit/Misfit Person Test

Entry Number	**Total Score**	**Total Count**	**Measure**	**Mode S.E**	**INFIT**		**OUTFIT**		**PT-MEASURE**		**EXACT MATCH**		**Person**
					MNSQ	**ZSTD**	**MNSQ**	**ZSTD**	**CORR.**	**EXP.**	**OBS%**	**EXP %**	
116	3	20	-2,07	0,69	0,99	0,1	5,01	3,0	0,14	0,38	90,0	86,8	116P
136	3	20	-2,07	0,69	0,81	-0,3	4,87	3,0	0,27	0,38	90,0	86,8	136P
112	4	20	-1,65	0,62	1,17	0,5	3,94	3,1	0,04	0,39	90,0	82,7	112P
135	4	20	-1,65	0,62	1,16	0,5	3,87	3,0	0,05	0,39	90,0	82,7	135L
140	4	20	-1,65	0,62	0,99	0,1	3,77	3,0	0,17	0,39	90,0	82,7	140L
160	4	20	-1,65	0,62	0,99	0,1	3,77	3,0	0,17	0,39	90,0	82,7	160P
170	4	20	-1,65	0,62	0,99	0,1	3,77	3,0	0,17	0,39	90,0	82,7	170P
133	4	20	-1,65	0,62	0,87	-0,2	3,43	2,7	0,30	0,39	90,0	82,7	133L
238	5	20	-1,30	0,57	1,18	0,7	2,89	2,8	0,11	0,40	75,0	79,1	238L
137	5	20	-1,30	0,57	0,87	-,03	2,59	2,4	0,36	0,40	85,0	79,1	137P
161	5	20	-1,30	0,57	0,87	-0,3	2,59	2,4	0,36	0,40	85,0	79,1	161P
159	5	20	-1,30	0,57	0,86	-0,4	2,58	2,4	0,36	0,40	85,0	79,1	159P
28	17	20	2,05	0,67	1,20	0,6	2,53	1,6	0,04	0,31	80,0	86,0	028P
227	7	20	-0,72	0,52	1,43	1,8	2,51	3,2	-0,09	0,41	70,0	71,2	227P
273	7	20	-0,72	0,52	1,40	1,7	2,45	3,1	-0,06	0,41	60,0	71,2	273L
56	11	20	0,26	0,49	1,25	1,6	2,31	3,2	0,03	0.39	55,0	65,7	056L
50	10	20	0,03	0,49	1,34	2,0	2,27	3,3	-0,02	0,40	60,0	66,0	050P
20	10	20	0,03	0,49	1,56	3,1	2,23	3,2	-0,17	0,40	40,0	66,0	020L
38	10	20	0,03	0,49	1,56	3,1	2,23	3,2	-0,17	0,40	40,0	66,0	038L
40	10	20	0,03	0,49	1,56	3,1	2,23	3,2	-0,17	0,40	40,0	66,0	040L
53	10	20	0,03	0,49	1,56	3,1	2,23	3,2	-0,17	0,40	40,0	66,0	053P
80	10	20	0,03	0,49	1,56	3,1	2,23	3,2	-0,17	0,40	40,0	66,0	080L
115	10	20	0,03	0,49	1,56	3,1	2,23	3,2	-0,17	0,40	40,0	66,0	115L
230	10	20	0,03	0,49	1,56	3,1	2,23	3,2	-0,17	0,40	40,0	66,0	230P
36	10	20	0,03	0,49	1,48	2,7	2,16	3,1	-0,11	0,40	50,0	66,0	036P
225	7	20	-0,72	0,52	1,21	1,0	2,12	2,5	0,12	0,41	60,0	71,2	225P
239	2	20	-2,62	0,81	1,45	0,9	2,03	1,1	-0,05	0,34	85,0	90,5	239P
236	3	20	-2,07	0,69	1,55	1,2	2,03	1,3	-1,0	0,38	80,0	86,8	236L
68	10	20	0,03	0,49	1,35	2,1	2,00	2,8	0,01	0,40	50,0	66,0	068L
46	16	20	1,65	0,60	1,00	0,1	1,90	1,3	0,22	0,33	85,0	81,9	046P
59	13	20	0,76	0,51	1,39	1,8	1,80	1,8	-0,03	0,37	60,0	69,7	059P
153	8	20	-0,46	0,50	1,09	0,5	1,76	2,1	0,23	0,40	70,0	68,9	153P
186	10	20	0,03	0,49	1,18	1,2	1,75	2,2	0,17	0,40	60,0	66,0	186P
17	14	20	1,03	0,53	1,26	1,1	1,74	1,5	0,07	0,36	70,0	73,6	017L
146	5	20	-1,30	0,57	1,27	0,9	1,72	1,4	0,10	0,40	75,0	79,1	146P
83	8	20	-0,46	0,50	1,41	2,0	1,66	1,9	0,02	0,40	50,0	68,9	083L
231	9	20	-0,21	0,49	1,15	0,9	1,65	2,0	0,19	0,40	70,0	67,4	231P
79	12	20	0,51	0,50	1,39	2,1	1,55	1,5	0,03	0,38	50,0	67,1	079P
MEAN	10,8	20,0	0,49	0,78	0,96	0,0	1,14	0,3			79,4	78,6	
S.D.	6,2	0,0	2,24	0,45	0,25	0,9	0,76	1,2			12,4	7,9	

As well as in the examination of the item, the suitability of the Rasch model on *person* or respondent also uses the same criteria (Boone et al., 2014). On these criteria, the ZSTD value is sample size's so influenced. If using large sample size, the ZSTD value will automatically increase, which is always above 3. Thus, if the sample size is below 500, it is not recommended to use the ZSTD value as the main reference (Sumintono & Widhiarso, 2015).

Based on Table 6. the fit/misfit person or students test results were presented above, of 283 students who took the test, 38 students (13.4%) declared as persons or respondents who had inconsistent answers (misfit). Of 38 students who were declared misfits, divided into 2 gender groups, are 23 female and 15 male. From these results, the declared misfit respondent suggested being eliminated from the analysis with inferential statistics. In addition, there are also 39 students (13.8%) who have maximum scores on the three criteria. So they are included in the upper outlier category. It is consistent with the results shown in Table.3. regarding the instrument statistics summary that states that the person or respondent *logit* value is -0,18.This logit value indicates that some students are in the outlier and misfit categories to the results are negative, it identifies a response pattern that is out of the ordinary.

CONCLUSION

Based on the research results that have been done, it can be said that: 1) In general, the Two-Tier Multiple Choice assessment instrument is able to explain 41.1% of the variance that appears in each group of respondents; 2) The validation results show that as many as 85% of the items are declared fit and 15% of the items are misfits, so improvements need to be made; 3) The results of student's ability, out of 283 people there are 13.4% have a misfit ability, 13.8% of students are included in the upper outliers ability group because they have a maximum score. And 72,8% of students have infit abilities. Therefore, this item is qualified to used as an assessment instrument in Integrated Science learning towards students abilities.

From the process of trials results carried out, and the conclusions that have been presented, it is necessary to make suggestions for further development. There are some proposed suggestions that is: 1) This assessment instrument was developed for junior high school students who focused on Integrated Science only. So it needs to be developed more for other learning materials and higher levels, such as high school and college; 2) In order to achieve the assessment instruments' product suitability, it is necessary to develop an assessment instrument with a larger and deeper scope. So that the results obtained are more maximal; and 3) The results of development research which the author is doing at this time there are still several factors that have not been fulfilled and become obstacles to the research process. So the further research is needed more on research materials and subjects so that the results obtained can also be maximized.

REFERENCE

Aprilia, N., Susilaningsih, E., Sudarmin, Sumarni, W., Mahatmanti, W. F., & Adhelia, N. U. (2021). *Aplikasi Model Rasch Pada Instrumen Tes Untuk Menganalisis Kemampuan Pemecahan Masalah Siswa Materi Larutan Asam Basa*. *13*(2), 106–118.

Bohori, M., & Liliawati, W. (2019). Analisis penguasaan konsep siswa menggunakan Rasch Model pada materi usaha dan energi. *Prosiding Seminar Nasional Fisika*, *0*.

Bond, T. (2015). *Applying the Rasch Model*. Routledge. https://doi.org/10.4324/9781315814698

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model : Fundamental Measurement in the HumBond, T. G., & Fox, C. M. (2007). Applying the Rasch Model : Fundamental Measurement in the Human Sciences Second Edition University of Toledo.an Sciences Second Edition University of Toledo*.

Boone, W. J., Yale, M. S., & Staver, J. R. (2014). Rasch analysis in the human sciences. In *Rasch Analysis in the Human Sciences*. https://doi.org/10.1007/978-94-007-6857-4

Djaali, H., Pudji M., Sudarmanto, Y. B. (2008). *Pengukuran dalam bidang pendidikan*.

Mardapi, D. (2016). *Pengukuran Penilaian dan Evaluasi Pendidikan*. Yogyakarta: Nuha Litera.

Mariam, S., Saleh, M., Warsono, W., & Mujiyanto, J. (2018). Using the Rasch Model for the Affective Assessment of EFL Learners. *Arab World English Journal*, *9*(2), 441–456. https://doi.org/10.24093/awej/vol9no2.29

Misbah, M., Mahtari, S., Wati, M., & Harto, M. (2019). Analysis of Students' Critical Thinking Skills in Dynamic Electrical Material. *Kasuari: Physics Education Journal (KPEJ)*, *1*(2), 103–110. https://doi.org/10.37891/kpej.v1i2.19

Rozeha, Azami, S. (2007). Application of Rasch Measurement in Evaluation of Learning Outcomes: A Case Study in Electrical Engineering. *Regional Conference on Engineering Mathematics, Mechanics, Manufacturing & Architecture*.

Rusmansyah, & Almubarak. (2020). *Students ' Cognitive Analysis Using Rasch Modeling As an Assessment for Planning of Strategies in Chemistry Learning*. *5*(3), 222–235.

Sihombing, R. U., Naga, D. S., & Rahayu, W. (2019). A Rasch Model Measurement Analysis on Science Literacy Test of. *Indonesian Journal Educational Review*, *6*(1), 44–55.

Sumintono, B., & Widhiarso, W. (2015). Aplikasi Pemodelan Rasch pada Assessment Pendidikan. In *Aplikasi Pemodelan Rasch pada Assessment Pendidikan* (p. 142).

Wigati, I., & Kurratul, A. (2020). *Analisis Penalaran Mahasiwa Pendidikan Biologi Pada Mata Kuliah Taksonomi Tumbuhan Dengan Model Rasch*. *12*(2), 145–153.

Winarti, A., & Mubarak, A. (2019). Rasch Modeling: A Multiple Choice Chemistry Test. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, *2*(1), 1–9. https://doi.org/10.23917/ijolae.v2i1.8985

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. https://doi.org/10.32698/0111