Integrated Science Education Journal
Vol. 5, No. 2, May 2024, pp. 101~109
ISSN: 2716-3725, DOI: 10.37251/isej.v5i2.864

Applying the Rasch Model to Assess Retention and Transfer Test Instruments in Science Education on Additive and Addictive Substances

Anisa Fitria[1], Jodion Siburian[1], Ilham Falani[1], Damris Muhammad[1,*]
[1]Master of Natural Sciences Education Study Program, Postgraduate, Universitas Jambi, Jambi, Indonesia

Article Info

Article history:

Received Dec 28, 2023
Revised Feb 12, 2024
Accepted Apr 20, 2024
OnlineFirst May 31, 2024

Keywords:

Rasch Model
Retention and Transfer Test Instrument
Science Subject

ABSTRACT

Purpose of the study: This study aims to evaluate the quality of items in retention and transfer tests related to additive and addictive substances. Using Rasch modeling, the study seeks to enhance the management of learning evaluations and improve our understanding of student abilities and question quality.

Methodology: The research utilizes the Rasch Model to analyze retention and transfer test instruments on science topics involving additives and addictive substances. Conducted with Winstep software, the analysis focuses on the performance of 92 purposively sampled 8th-grade students during their first semester of junior high school. The study examines retention and transfer abilities, comprehensively evaluating the test items.

Main Findings: The Winstep program analysis reveals that, according to the Rasch model, the average ± MNSQ Outfit values for both items and persons are 0.92. The Outfit ZSTD values for items and persons are -0.12 and -0.01, respectively. The instrument's reliability, measured by Cronbach's alpha, is 0.60, indicating moderate reliability. The research findings demonstrate that each item in the instrument is valid and reasonably reliable, with all 20 items deemed suitable for assessing student performance in retention and transfer tests.

Novelty/Originality of this study: This study offers a detailed examination of retention and transfer test instruments' quality using the Rasch Model, providing valuable insights for enhancing the accuracy and reliability of these assessment tools. The research significantly improves educational assessments in science education, particularly in evaluating students' understanding of additives and addictive substances.

Corresponding Author:
Damris Muhammad,
Master of Natural Sciences Education Study Program, Postgraduate Program, Universitas Jambi
Jl. Lintas Jambi-Muaro Bulian. Muaro Jambi, Jambi, Indonesia.
Email: damris@unja.ac.id

1. INTRODUCTION

Science education, especially in the realm of Natural Sciences (IPA), plays a crucial role in shaping students' understanding of scientific concepts. One of the in-depth topics in science learning is the topic of additives and addictive substances. Good understanding of this material not only requires students to remember the information (retention test), but also to be able to apply these concepts in new contexts (transfer test). Students' competency in mastering science material and concepts is still considered inadequate [1]-[4]. Although there are various evaluation methods used in science education, many have not been able to effectively measure students' understanding and transfer ability in additive and additive materials. Previous research shows that there are still deficiencies in the evaluation instruments used, both in terms of validity and reliability [5], [6]. This indicates an urgent need to develop better evaluation instruments.

In an effort to improve the quality of measuring student performance in science material on additives and addictive substances, an evaluation instrument is needed that can provide valid and reliable results. Evaluation is an assessment action that aims to measure the level of success of a learning process [7], [8]. In measurement, there is a comparison between a prediction and a certain standard. The results of this measurement are then interpreted and considered as part of the assessment process [9], [10]. The use of the Rasch Model, as a measurement analysis method in the context of tests, is becoming increasingly important to ensure that the instruments developed have the right level of difficulty and can provide accurate information about students' abilities [11], [12]

The Rasch model is one of the modern analytical models that can be used to determine the feasibility of an instrument [4]. The Rasch model can be used to assess the level of index reliability [13], analyze items at each level [14], evaluate respondent reliability [15], as well as detect bias for each item on the test instrument and identify dimensionality [16]. Another advantage of the Rasch model is its capability to accommodate tests with item administration designs that involve random distribution [17], [18]

Through the implementation of the Rasch Model in the development of two-tier test instruments (retention tests and transfer tests), it is hoped that an evaluation instrument that is adequate and sensitive to differences in student abilities can be created. This will support teachers in providing more accurate feedback to improve the learning process and students' understanding of additive and addictive substance material.

Question item analysis can be used as diagnostic information to determine whether students understand what they have learned and to improve the quality of questions by correcting or eliminating invalid questions [19]. This research aims to detect the quality of the items in the assessment of retention tests and transfer tests on additive and addictive substance material. which is useful for improving the management of the implementation of learning evaluations and increasing information about students' abilities and the quality of the questions given, analysis of the questions is carried out using Rasch modeling.

2. RESEARCH METHOD

This research is a quantitative - descriptive study. This research applies the Rasch Model to evaluate retention test instruments and transfer tests on additive and addictive science material. This analysis was carried out with the help of Winsteps software. The data collection method was carried out through instrument sheets in the form of retention tests and transfer tests, with a focus on the responses of Class VIII students at Junior high school 7 Muaro Jambi when taking retention tests and transfer tests. The data used in this research is dichotomous, with a dichotomous scoring model which refers to the dichotomous logistic model, in accordance with the principles explained by Hambleton, Swaminathan, and Rogers in Retnawati [20]. The population in this study consists of Class VIII students at Junior high school 7 Muaro Jambi. The sampling technique used is purposive sampling, where the sample is selected based on specific criteria relevant to the research objectives. The number of samples taken is 75 students.

The instruments used in this research are retention tests and transfer tests, each consisting of 20 questions, focusing on science material on additives and addictive substances. These tests are designed to measure student performance in understanding science learning through multimedia. These instruments are adapted from previous research that has been tested for validity and reliability. In this study, the instruments were adopted and adapted according to the context and needs of the research. Data collection was conducted through retention tests and transfer tests administered to the students. The data collected consists of student response data using multiple choice questions that are dichotomous, i.e., correct or incorrect.

Data analysis was carried out using the Rasch Model, implemented through Winsteps software [24]. The Rasch Model is used to develop a measurement model that relates the respondent's ability level (person ability) to the difficulty level of the items (item difficulty). This analysis includes:

1. Reliability: Produces values for item reliability, individual reliability, and Cronbach's alpha (item-person).
2. Dimensionality: Conducted to determine whether the measurement instrument focuses on only one dimension. An optimal instrument is unidimensional, meaning it measures only one variable or concept [21].
3. Item and Respondent Fit: Analysis of item or respondent parameters is used to assess the fit between items and responses with the model. Fit criteria include Point Measure Correlation ($0.32 < x < 0.8$), Outfit Mean Square ($0.5 < y < 1.5$), and Outfit Z-standard ($-2.0 < z < +2.0$). Data that fits the Rasch model has a mean square value of 1.0 and a Z-standard value of 0.0.
4. DIF Analysis: Conducted to assess whether the items compiled contain bias or favor one group, such as gender. If the probability value (PROB) in the table is less than 0.05, it indicates that the item contains bias.

By using the Rasch Model, this research aims to evaluate the quality of the instruments and adjust the parameters to suit the characteristics of the respondents and the desired measurement objectiv. The Rasch model

develops a measurement model that can relate the relationship between the level of student respondent's ability (person ability) and the level of item difficulty (item difficulty). Student respondents with high skills will be able to solve questions that have a lower level of difficulty [22]. The Rasch model can analyze data that is dichotomous or polytomous. The Rasch model assumes that the level of difficulty of an item is influenced by the response from the respondent, while individual ability is influenced by the estimated level of difficulty of an item [23]

The assessment of the retention test and transfer test instruments in this study includes research on reliability, dimensions, validity, and differential item function (DIF) analysis. The data analysis processing process was carried out using the Winstep software application [24]. Reliability analysis using the Rasch model produces item reliability values, individual reliability, and Cronbach's alpha (item-person). Dimensionality analysis is carried out to determine whether the measurement instrument only focuses on one dimension. The optimal instrument is unidimensional, measuring only one variable or concept. For can be considered a unidimensional instrument, an instrument must meet the requirements of having a minimum raw variance value of 20%.

Analysis of item or respondent parameters used to assess the suitability between items and responses with the model must meet three criteria, namely: Point Measure Correlation (x) : $0.32 < x < 0.8$ then Outfit Mean Square (y): $0.5 < y < 1.5$, and Outfit Z standard(z): $-2.0 < z < +2.0$. Data that corresponds to the Rasch model has a mean square value of 1.0 and a Z-standardized value of 0.0. An item is said to be misfit if the item is too easy or in other words the logit value is too negative or too difficult (a large positive logit value). Or, if the resulting logit value does not meet the three criteria mentioned above, items that not meeting these requirements indicates that the item does not adequately measure the desired trait or trait [30].

DIF analysis aims to assess whether the items compiled contain bias or favor one party, such as gender. The criteria applied is to look at the PROB value in the table. if the probability value is less than 0.05, it indicates that the item contains bias. In this research, DIF analysis was carried out based on the gender variable.

Table 1. Categories of Reliability-Person Values and Reliability-Item Items

Person-Reliability and Item-Reliability Values	Category
< 0.67	Weak Category
0.67 – 0.80	Sufficient Category
0.80 – 0.90	Good Category
0.91 – 0.94	Very Good Category
> 0.94	Special Category

Table 2. Cronchbach Alpha Value Criteria (Reliability) for Question Items

Alpha-Cronchbach Value (Reliability)	Criteria
< 0.50	Bad
0.50 – 0.60	Bad
0.60 – 0.70	Enough
0 .70 – 0.80	Good
> 0.80	Very good

Calculation of Item Fit Order The level of suitability of the items (validity) which is used to explain whether each item operates normally in carrying out measurements. Item Measure, in this context, reflects the level of difficulty of each question item.

Table 3. Difficulty Level Categories of Question Items

Logit Value	Category
Greater than +1.37SD	Very difficult
0.0 Logit +1.37SD	Difficult
0.0 Logit -1.37SD	Currently
Smaller than -1.37SD	Easy

In contrast to the Item Measure which is the level of difficulty of the question item, there is a Person Measure which is the level of student ability or the student's ability to answer questions.

Table 4. Criteria for Grouping Student Abilities

Logit Value of Student Ability	Criteria
Greater than 1.80	Tall
Less than 1.80	Currently
Less than -1.29	Low

Next, the Person fit order is used to detect individuals whose response patterns are inappropriate or different. Based on [24], Analysis of item or respondent parameters used to assess the suitability between items and responses with the model must meet three criteria, namely: Point Measure Correlation (x): $0.32 < x < 0.8$ then Outfit Mean Square (y) : $0.5 < y < 1.5$, and Outfit Z standard(z): $-2.0 < z < +2.0$. An item is considered misfit if it is too easy (with a very negative logit value) or too difficult (with a very positive logit value); or if the resulting logit value does not meet the three criteria mentioned above. Items that do not meet these requirements indicate that the item does not adequately measure the desired characteristics or traits [30]. Apart from that, the Rasch model used can also analyze Differential Item Functioning (DIF) [25]. The DIF analysis itself was carried out in this study based on gender. The criteria used in the analysis are probability values, if the prob <0.05 means the item is categorized as containing bias (DIF).

3. RESULTS AND DISCUSSION

Retention tests and transfer tests are two tier standard tests that have the ability to measure the ability to remember and understand material. This test is used to measure the extent of student performance through the application of multimedia. The retention test and transfer test consist of 10 multiple choice questions each for retention questions and 10 multiple choice questions for transfer test questions with a total of 20 questions.

The Rasch model has the ability to carry out analysis of test items, respondent analysis, and even achieve comprehensive instrument analysis [26]. Another advantage of rash analysis lies in its ability to provide statistical summary results and very detailed test information functions [27]. The information function provided by the Rasch model is comprehensive, providing guidance to instrument makers to make logical, precise and scientific decisions based on in-depth analysis. In testing the quality of the instrument, the Rasch model is used with the aim of providing comprehensive information about the quality of the response patterns of student respondents as a whole, the quality of the test instruments, and the interaction between student respondents and the research instruments used. The results of this analysis can be found in Figure 1.

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	13.6	20.0	.99	.56	1.00	.10	.92	-.01
SEM	.3	.0	.09	.01	.03	.11	.05	.11
P.SD	2.8	.0	.80	.07	.23	.94	.39	.93
S.SD	2.8	.0	.81	.07	.24	.94	.39	.94
MAX.	18.0	20.0	2.58	.78	1.59	2.34	2.44	2.36
MIN.	6.0	20.0	-1.04	.50	.60	-2.15	.34	-1.96

REAL RMSE	.59	TRUE SD	.55	SEPARATION	.94	Person RELIABILITY	.47
MODEL RMSE	.56	TRUE SD	.57	SEPARATION	1.01	Person RELIABILITY	.51
S.E. OF Person MEAN = .09							

Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .60 SEM = 1.87
STANDARDIZED (50 ITEM) RELIABILITY = .72

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	51.0	75.0	.00	.30	1.00	.19	.92	-.12
SEM	3.0	.0	.23	.01	.03	.24	.05	.23
P.SD	13.3	.0	1.01	.05	.13	1.05	.20	1.00
S.SD	13.6	.0	1.04	.06	.13	1.08	.20	1.03
MAX.	69.0	75.0	1.93	.44	1.24	2.39	1.29	1.69
MIN.	23.0	75.0	-1.72	.25	.78	-1.67	.60	-1.46

REAL RMSE	.31	TRUE SD	.96	SEPARATION	3.13	Item RELIABILITY	.91
MODEL RMSE	.30	TRUE SD	.97	SEPARATION	3.20	Item RELIABILITY	.91
S.E. OF Item MEAN = .23							

Item RAW SCORE-TO-MEASURE CORRELATION = -.99
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000

Figure 1. Reliability value of retention test instruments and transfer tests on additive and addictive substances

Based on the output image of the Winstep application table above, which is the result of data analysis of student responses, it can be seen that the average score of respondents in working on retention and transfer test questions by student respondents is 0.99. This means that the average value obtained is greater than the logit value of 0.00, which indicates that the ability of the student respondents is lower than the difficulty level of the questions.

In the results of the analysis of the reliability of the retention test and transfer test instruments on additive and addictive substance material, it was obtained at 0.91 and the person-reliability was obtained at 0.51. This shows that the reliability of the retention test instrument and transfer test instrument for additive and Addictive is included in the good category. In the Rasch model, apart from being able to obtain the results of the reliability analysis of instruments and people, you can also obtain Cronbach-Alpha values, which is the reliability of the interaction between the person and the items as a whole. Based on the results of the analysis, it can also be seen that the Cronbach's Alpha value obtained was 0.60, which is included in the sufficient category. The results of the analysis can be seen in table 1.

Next, the average values of the OUTFIT MNSQ and INFIT MNSQ person are 0.92 and 1.00 respectively. These results indicate that the values obtained are at a level that is considered ideal. In addition, for the INFIT - ZSTD and OUTFIT ZSTD values, it was found to be 0.10 and -0.01, which are close to the ideal value of 0.0. This shows that the quality is getting better. The level of item difficulty is a parameter that describes how difficult it is for a group of respondents taking retention tests and transfer tests on additive and addictive substance material to provide the correct response to an item. Information about item difficulty levels can be found in figure 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE 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
15	23	75	1.93	.26	1.24	2.05	1.29	1.65	.03	.31	62.7	71.5	P15
13	31	75	1.40	.25	1.20	2.39	1.20	1.63	.10	.34	57.3	65.0	P13
14	32	75	1.34	.25	1.01	.21	.97	-.26	.34	.34	61.3	64.3	P14
9	38	75	.98	.25	.88	-1.67	.86	-1.46	.49	.35	68.0	64.3	P9
3	40	75	.86	.25	.98	-.23	.94	-.53	.38	.35	65.3	64.8	P3
12	41	75	.79	.25	1.03	.38	1.10	.95	.30	.35	69.3	65.0	P12
10	44	75	.61	.25	1.17	1.85	1.20	1.69	.15	.35	54.7	66.3	P10
2	49	75	.29	.26	1.13	1.27	1.14	.99	.19	.35	62.7	69.7	P2
11	49	75	.29	.26	1.01	.13	.94	-.41	.36	.35	65.3	69.7	P11
4	50	75	.22	.26	.90	-.87	.90	-.68	.45	.34	73.3	70.6	P4
1	51	75	.15	.26	1.11	1.00	1.11	.78	.21	.34	66.7	71.5	P1
17	59	75	-.47	.30	1.08	.50	.97	-.03	.26	.32	77.3	79.8	P17
6	60	75	-.56	.30	.96	-.18	.79	-.81	.40	.31	80.0	80.8	P6
7	60	75	-.56	.30	.83	-.97	.73	-1.12	.51	.31	85.3	80.8	P7
8	63	75	-.86	.33	.78	-1.05	.62	-1.37	.55	.29	86.7	84.2	P8
16	64	75	-.98	.34	.95	-.15	.79	-.61	.37	.29	85.3	85.4	P16
5	66	75	-1.23	.37	.90	-.30	.65	-.95	.42	.27	88.0	88.0	P5
18	66	75	-1.23	.37	1.04	.23	.85	-.30	.26	.27	88.0	88.0	P18
20	66	75	-1.23	.37	.82	-.67	.60	-1.11	.50	.27	88.0	88.0	P20
19	69	75	-1.72	.44	.93	-.09	.73	-.43	.33	.23	92.0	92.0	P19
MEAN	51.0	75.0	.00	.30	1.00	.19	.92	-.12			73.9	75.5	
P.SD	13.3	.0	1.01	.05	.13	1.05	.20	1.00			11.6	9.4	

Figure 2. Analysis of the difficulty level of the questions

Based on the analysis of the level of difficulty of the questions, it can be seen that question number 15 (P15) in the retention test and transfer test of additive and addictive material has the highest logit value, namely 1.93, indicating that this question is the most difficult. Followed by questions number 13 and 14 (P13) and (P14) which are also included in the most difficult question category. On the other hand, question number 19 (P19) is considered the easiest question with a logit value of -1.72. According to the Rasch model, the higher the logit value, the greater the level of difficulty of the question, and vice versa, if the lower the logit value, the easier it will be for students to answer the question. The results of the analysis of the level of difficulty of the questions using the Winstep application provide more convenience to users because the results of the analysis of the level of difficulty of the questions have been sorted from the highest level of difficulty in this case question number 15 (P15) to the lowest level of difficulty in this case question number 19 (P19). Of course, the results of this

analysis make it easier to identify questions that have a high level of difficulty and questions that have a low level of difficulty.

In the analysis using the Rasch model, apart from the level of difficulty of the question items, the quality of the suitability of the question items with the model, namely the item fit model, can also be produced. Item fit analysis is an evaluation of whether an item functions normally in the measurement process. If there is a mismatch (fit), this can indicate potential misconceptions experienced by the respondent regarding the question item. In other words, the results of item fit analysis provide an overview of the extent to which the items can provide an appropriate measure of the characteristics or skills being measured, as well as the extent to which the respondents' responses are in line with expected expectations. This analysis plays a critical role in ensuring the integrity and validity of the measurement instrument. Based on research by Boone et al. (2014) and Bond and Fox (2015), criteria such as outfit mean square value, point measure correlation, and outfit z-standard are used to assess the level of suitability of the question items (item fit). If a question item does not meet all three of these criteria, it can be concluded that the quality of the question item can be said to be poor and needs to be revised or even replaced. This is necessary to ensure that the respondent's understanding is truly tested through the use of appropriate and good quality question items. The item fit indicators for all question items involve 3 things that need to be fulfilled, namely Outfit Z Standard (-2.0 <ZSTD < +2.0), Outfit Mean Square (0.5 < MNSQ < 1.5), and Pt Measure Correlation (0.4 < Pt Measure Corr < 0.85). The results of the analysis of the suitability of retention test items and transfer tests on additive and addictive substances can be found in Figure 3.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE 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
15	23	75	1.93	.26	1.24	2.05	1.29	1.65	A .03	.31	62.7	71.5	P15
10	44	75	.61	.25	1.17	1.85	1.20	1.69	B .15	.35	54.7	66.3	P10
13	31	75	1.40	.25	1.20	2.39	1.20	1.63	C .10	.34	57.3	65.0	P13
2	49	75	.29	.26	1.13	1.27	1.14	.99	D .19	.35	62.7	69.7	P2
1	51	75	.15	.26	1.11	1.00	1.11	.78	E .21	.34	66.7	71.5	P1
12	41	75	.79	.25	1.03	.38	1.10	.95	F .30	.35	69.3	65.0	P12
17	59	75	-.47	.30	1.08	.50	.97	-.03	G .26	.32	77.3	79.8	P17
18	66	75	-1.23	.37	1.04	.23	.85	-.30	H .26	.27	88.0	88.0	P18
11	49	75	.29	.26	1.01	.13	.94	-.41	I .36	.35	65.3	69.7	P11
14	32	75	1.34	.25	1.01	.21	.97	-.26	J .34	.34	61.3	64.3	P14
3	40	75	.86	.25	.98	-.23	.94	-.53	j .38	.35	65.3	64.8	P3
6	60	75	-.56	.30	.96	-.18	.79	-.81	i .40	.31	80.0	80.8	P6
16	64	75	-.98	.34	.95	-.15	.79	-.61	h .37	.29	85.3	85.4	P16
19	69	75	-1.72	.44	.93	-.09	.73	-.43	g .33	.23	92.0	92.0	P19
4	50	75	.22	.26	.90	-.87	.90	-.68	f .45	.34	73.3	70.6	P4
5	66	75	-1.23	.37	.90	-.30	.65	-.95	e .42	.27	88.0	88.0	P5
9	38	75	.98	.25	.88	-1.67	.86	-1.46	d .49	.35	68.0	64.3	P9
7	60	75	-.56	.30	.83	-.97	.73	-1.12	c .51	.31	85.3	80.8	P7
20	66	75	-1.23	.37	.82	-.67	.60	-1.11	b .50	.27	88.0	88.0	P20
8	63	75	-.86	.33	.78	-1.05	.62	-1.37	a .55	.29	86.7	84.2	P8
MEAN	51.0	75.0	.00	.30	1.00	.19	.92	-.12			73.9	75.5	
P.SD	13.3	.0	1.01	.05	.13	1.05	.20	1.00			11.6	9.4	

Figure 3. Level of suitability of question items

Based on the results of the table output from the Winstep application as in Figure 3 above regarding the analysis of the level of suitability of the question items, it can be seen that in the items there are no questions that do not meet the three criteria, each question item has an average of 2 of the 3 criteria. so there is no need to throw away any questions. However, there are 13 items that must be revised, namely items 15 (P15), 13 (P13), 14 (P14), 3 (P3), 12 (P12), 10 (P10), 2 (P2), 11 (P11), 1 (P1), 17 (P17), 16 (P16), 18 (P18), AND 19 (P19).

One indicator of the validity of a measurement is the absence of bias in the instrument and each question item used [30]. An instrument or question item is said to have bias when it is found that one individual with certain characteristics gets greater benefits compared to other individuals who have other characteristics. In the Rasch model, bias identification is known as differential item functioning (DIF). In this study, DIF analysis was carried out to assess whether the items compiled contained bias based on gender or not. The table is used as a criterion, where a probability value (PROB) of less than 0.05 indicates the presence of bias in an item. The results of the DIF analysis can be accessed in Figure 4.

Person CLASSES	SUMMARY DIF CHI-SQUARED	D.F.	PROB.	BETWEEN-CLASS/GROUP Item UNWTD MNSQ	ZSTD	Number	Name
2	.8717	1	.3505	.8947	.39	1	P1
2	.2562	1	.6127	.2620	-.29	2	P2
2	2.8562	1	.0910	3.0050	1.41	3	P3
2	.0363	1	.8488	.0367	-.94	4	P4
2	.0907	1	.7633	.0923	-.69	5	P5
2	.2032	1	.6521	.2070	-.40	6	P6
2	.2032	1	.6521	.2070	-.40	7	P7
2	.1421	1	.7062	.1446	-.54	8	P8
2	.4614	1	.4970	.4705	.00	9	P9
2	.2576	1	.6117	.2620	-.29	10	P10
2	1.1285	1	.2881	1.1631	.58	11	P11
2	.8303	1	.3622	.8527	.36	12	P12
2	2.0547	1	.1517	2.1409	1.08	13	P13
2	.4071	1	.5235	.4146	-.07	14	P14
2	.5001	1	.4795	.5108	.05	15	P15
2	2.1644	1	.1412	2.3255	1.16	16	P16
2	.0038	1	.9509	.0053	-1.28	17	P17
2	.0907	1	.7633	.0923	-.69	18	P18
2	.0495	1	.8240	.0503	-.87	19	P19
2	1.0851	1	.2976	1.1331	.56	20	P20

Figure 4. DIF value for each item on the retention test instrument and transfer test on additive and addictive substance material

Based on the Winstep output table provided, it can be concluded that there are no items that show DIF (Differential Item Functioning). This means that the twenty items are not specifically beneficial or detrimental for only one gender. This DIF analysis shows that there are no systematic differences in the way male student respondents and female student respondents answer questions, so that this item can be considered to have no bias influence towards certain gender groups.

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = Item information units

	Eigenvalue	Observed		Expected
Total raw variance in observations =	25.8517	100.0%		100.0%
Raw variance explained by measures =	5.8517	22.6%		23.9%
Raw variance explained by persons =	2.3844	9.2%		9.8%
Raw Variance explained by items =	3.4674	13.4%		14.2%
Raw unexplained variance (total) =	20.0000	77.4%	100.0%	76.1%
Unexplned variance in 1st contrast =	3.0502	11.8%	15.3%	
Unexplned variance in 2nd contrast =	2.2243	8.6%	11.1%	
Unexplned variance in 3rd contrast =	2.1340	8.3%	10.7%	
Unexplned variance in 4th contrast =	1.5837	6.1%	7.9%	
Unexplned variance in 5th contrast =	1.3906	5.4%	7.0%	

Figure 5. Raw Variance Value of retention test instruments and transfer tests on additive and addictive substances

Based on the Rasch Model analysis, it was found that the set of retention and transfer test instruments on additive and addictive substances had a Raw Variance value of 23.9%. These results indicate that the test instrument has good construct validity. Raw variance is considered a latent measure of an instrument being measured, and this helps verify the measurement construction of the instrument. Validity here is not only related to the content, but also includes the consequences of using test scores. By using the Rasch measurement model, analysis can assess the extent to which test items correspond to the identified constructs of the test instrument as a whole. The results of this analysis provide information about the suitability of the items to the general construction of the test instrument.

Each measurement always provides data regarding the measurement results, and the measurement information obtained is very dependent on the relationship between the individual and the test being measured.

The information resulting from measurements is influenced by the degree of variation in the results obtained. In the framework of the theory of item response according to the model, the information function itself can be divided into two aspects, namely the item information function for each item and also the test information function for the entire test. In item response theory, the information function of the item is the main determinant in selecting appropriate items, in contrast to the item analysis approach in classical test theory. In addition, to evaluate the overall quality of a test consisting of a number of items, item response theory tends to utilize the information function of the test rather than relying on reliability coefficients, which is the main approach in classical test theory.

The information function is used to describe the strength of the question items on the instrument or test, select instrument or test items and compare test instruments [28], [29] The information function indicates the extent of superiority or contribution of a test in describing the hidden characteristics measured by the test. When the skill level is low, the level of information resulting from the measurement is also low. Likewise, the level of ability or ability of respondents is very high and medium. At a medium level of capability, the information obtained by measurement is very high. This shows that retention test items and transfer tests on additive and addictive substances produce optimal information when given to subjects with moderate abilities. This means that the test questions are tests with a medium level of difficulty [30]. This is supported by the results of the test reliability coefficient analysis in Figure 1, namely 0.99, where the measure or logit value is = 0, indicating that the retention and transfer tests suitable for measuring student performance after being exposed to multimedia.

4. CONCLUSION

Based on the research and discussions that have been carried out, several conclusions can be drawn. The results of the reliability analysis of the retention test and transfer test instruments on additives and addictive substances showed a value of 0.99, while individual reliability (person reliability) was obtained at 0.51. This shows that the reliability of retention tests and transfer tests on science material with additives and addictive substances is in the sufficient category. Based on the results of the analysis, it can be seen that the Cronbach's Alpha value obtained at 0.60 is included in the sufficient category.Based on the results of the analysis, it can also be seen that the Cronbach's Alpha value obtained was 0.60, which is included in the sufficient category.

Based on the results of data analysis regarding the level of difficulty of the questions, it can be seen that question number 15 (P15) in the retention test and transfer test of science material on additives and addictive substances has the highest logit value, namely 1.93, indicating that this question is the most difficult. Followed by questions number 13 and 14 (P13) and (P14) which are also included in the most difficult question category. On the other hand, question number 19 (P19) is considered the easiest question with a logit value of -1.72. The average INFIT MNSQ and OUTFIT MNSQ for individuals were 1.00 and 0.92 respectively, indicating that the values obtained were at a level considered ideal. Meanwhile, the INFIT ZSTD and OUTFIT ZSTD values were found to be 0.10 and -0.01 respectively, close to the ideal value of 0.0, which indicates an increase in quality. Likewise, the mean INFIT MNSQ and OUTFIT MNSQ for the items were 1.00 and -0.92 respectively, indicating the resulting values were in line with the desired standards. For INFIT ZSTD and OUTFIT ZSTD, the figures are 1.0 and -0.1, close to the ideal value of 0.0, indicating increased quality.

Based on the analysis of the level of suitability of the question items, it can be seen that in the items there are no questions that do not meet the 3 criteria, each question item on average has 2 of the 3 criteria. so there is no need to throw away any questions. However, there are 13 items that need to be revised. Based on the DIF analysis, we can see that there are no items that show DIF (Differential Item Functioning). This DIF analysis shows that there are no systematic differences in the way male and female student respondents answer questions, so that this item can be considered to have no biasing influence towards certain gender groups.

ACKNOWLEDGEMENTS

The researcher would like to thank all parties involved in the research conducted by the researcher.

REFERENCES

[1] U. Usmeldi, R. Amini, "The effect of integrated learning model to the students competency on the natural science," *J Phys Conf Ser*, 2019, doi: 10.1088/1742-6596/1157/2/022022.
[2] J. Suroso, I, Indrawati, S. Sutarto, I. Mudakir, "Profile of high school students science literacy in east java," *J Phys Conf Ser*, 2021, doi: 10.1088/1742-6596/1832/1/012040.
[3] A. K. Dewi, "Improving Students Learning Outcomes Through Mind Map in Human Reproductive System Topic in Natural Science Learning," *Int J Educ Vocat Stud*, vol. 1, pp. 702-706, 2019, doi: 10.29103/ijevs.v1i7.1675.
[4] S. Maulana, A. Rusilowati, S. E. Nugroho, E. Susilaningsih, "Implementasi Rasch Model dalam Pengembangan Instrumen Tes Diagnostik," *Pros Semin Nas Pascasarj*, vol. 6, pp. 748-756, 2023.
[5] Md. Ghazali, "A Reliability and Validity of an Instrument to Evaluate the School-Based Assessment System: A Pilot Study," *Int J Eval Res Educ*, vol. 5, pp. 148, 2016, doi: 10.11591/ijere.v5i2.4533.

[6] A. Maulana, "Analysis of Validity, Reliability and Feasibility of Student Confidence Assessment Instruments," *Schola*, vol. 1, pp. 1-12, 2023.
[7] E. Romiyati, A. Ardi Rahman, and E. Budiyono, "Development of Mathematical Student Worksheets Based on Scientific Approaches and PQ4R Learning Strategies on Associated Materials", *Jor. Eva. Edu*, vol. 4, no. 1, pp. 17-20, 2023, doi: 10.37251/jee.v4i1.296.
[8] A. Doelvia, V. T. T. Hien, and S. Rathee, "Assessment: The Effectiveness of Video Media Through the Tiktok Application on Teenagers' Knowledge About Clean and Healthy Living Behavior at Junior High School Level", *Jor. Eva. Edu*, vol. 4, no. 4, pp. 168-174, 2023, doi: 10.37251/jee.v4i4.948.
[9] Z. Zainal, "Pengukuran, Assessment dan Evaluasi dalam Pembelajaran Matematika," *J Pendidik Mat*, vol. 3, pp. 8-26, 2020.
[10] V. P. Sabandar, and H. B. Santoso, "Evaluasi Aplikasi Media Pembelajaran Statistika Dasar Menggunakan Metode Usability Testing," *Teknika*, vol. 7, pp. 50-59, 2018, doi: 10.34148/teknika.v7i1.81.
[11] M. Ibnu, B. Indriyani, H. Inayatullah, and Y. Guntara, "Aplikasi rasch model: Pengembangan instrumen tes untuk mengukur miskonsepsi mahasiswa," *Pros Semin Nas Pendidik FKIP*, vol. 2, pp, 205-210, 2019.
[12] S. E. Mokshein, H. Ishak, H. Ahmad, "The use of rasch measurement model in English testing," *Cakrawala Pendidik*, vol. 38, pp. 16-32, 2019, doi: 10.21831/cp.v38i1.22750.
[13] N. Ngadi, "Analisis model rasch untuk mengukur kompetensi pengetahuan siswa smkn 1 Kalianget pada mata pelajaran perawatan sistem kelistrikan sepeda motor," *J Pendidik Vokasi Otomotif*, vol. 6, pp. 1-19, 2023.
[14] G. Hamdu, F. N. Fuadi, A. Yulianto, Y. S. Akhirani, "Items Quality Analysis Using Rasch Model To Measure Elementary School Students' Critical Thinking Skill On Stem Learning," *JPI (Jurnal Pendidik Indones)*, vol. 9, no. 1, 2020, doi:.org/10.23887/jpi-undiksha.v9i1.20884.
[15] T. Rachman, D. B. Napitupulu, "Rasch model for validation a user acceptance instrument for evaluating e-learning system," *CommIT (Communication Inf Technol J)*, vol. 11, no. 1, 2017, doi: 10.21512/commit.v11i1.2042.
[16] Y. E. Suryani, "Aplikasi rasch model dalam mengevaluasi intelligenz structure test (IST)," *Psikohumaniora J Penelit Psikol*, vol. 3, no. 1, pp. 73, 2018, doi: 10.21580/pjpp.v3i1.2052.
[17] W. Widhiarso, "Penerapan model rasch untuk mengevaluasi Tes UKKS dan UKPS," *Tenaga Kependidikan*, vol. 1, pp. 50-61, 2016.
[18] D. Pratama, I. Husnayaini, "Applying rasch model to measure students` reading comprehension, "*JISAE J Indones Student Assess Eval*, vol. 6, no. 2, pp. 203-209, 2020, doi: 10.21009/jisae.v6i2.14920.
[19] A. Fauziana, D. Wulansari, "Analisis kualitas butir soal ulangan harian di sekolah dasar dengan model rasch," *Ibriez J Kependidikan Dasar Islam Berbas Sains*, vol. 6, no. 1, pp. 10-19, 2021, doi: 10.21154//ibriez.v6i1.112.
[20] H. Retnawati, *Teori Respon Butir dan Penerapanya*, Yogyakarta Nuha Med, 2014.
[21] R. Tate, "Test dimensionality," In *Large-scale assessment programs for all students* (pp. 155-180). Routledge, 2012.
[22] F. Bond, "Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.)," Lawrence Erlbaum Associates Publishers, 2007.
[23] J. M. Linacre, "Understanding rasch measurement: Optimizing rating scale category effectiveness," *J Appl Meas*, vol. 3, pp. 85-106, 2002.
[24] W. J. Boone, R. J. Staver, and S. M. Yale, *Rasch analysis in the human sciences*. London: Springer; 2014.
[25] S. Sudaryono, "Sensitivity of differential item functioning (dif) Detection Method," *J Eval Pendidik*, vol. 3, pp. 82-94, 2012.
[26] A. Darmana, A. Sutiani, H. A. Nasution, I. Ismanisa, N. Nurhaswinda, "Analysis of rasch model for the validation of chemistry national exam instruments," *J Pendidik Sains Indones*, vol. 9, no. 3, pp. 329-345, 2021, doi: 10.24815/jpsi.v9i3.19618.
[27] L. Tesio, A. Caronni, D. Kumbhare, S. Scarano, "Interpreting results from Rasch analysis 1. The "most likely" measures coming from the model," *Disabil Rehabil*, vol. 46, pp. 591-603, 2024 doi: 10.1080/09638288.2023.2169771.
[28] H. Hambleton, *Fundamentals of item response theory*, Sage Publications, Inc. 1991.
[29] U. D. Purnamasari, B. Kartowagiran, "Application rasch model using R program in analyze the characteristics of chemical items," *J Inov Pendidik IPA*, vol. 5, no. 2, pp. 147-157, 2019, doi: 10.21831/jipi.v5i2.24235.
[30] B. Sumintono, and W. Widhiarso, *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. Cimahi: Trim Komunikata Publishing House. 2015.