Jurnal Penelitian dan Evaluasi Pendidikan
Volume 27, No 1, June 2023 (76-91)
Online: http://journal.uny.ac.id/index.php/jpep

Item quality analysis using the Rasch model to measure critical thinking ability in the material of the human digestive system of Biology subject in high school

Wanda Agus Prasetya*; Anggi Tias Pratama
Universitas Negeri Yogyakarta, Indonesia
*Corresponding Author. E-mail: wandaagus.2021@student.uny.ac.id

ARTICLE INFO

Article History
Submitted:
23 February 2023
Revised:
6 June 2023
Accepted:
6 June 2023

Keywords
Rasch model; critical thinking ability; human digestive system

Scan Me:

ABSTRACT

This study aims to determine the quality of the Biology instrument items on the digestive system material with the Rasch model for analyzing critical thinking skills. The research employed a quantitative descriptive method involving 63 students of senior high schools in Yogyakarta. The data were collected using a critical thinking skills description test and processed using the Rasch model with the Winstep 5.0.3.4 program. The study shows that the overall validity is acceptable. The item validity did not require improvement in items 14, 3, 8, 13, 12, 1, 4, 10, 6, 2, 7, 11, 15, and 9, and required improvement or replaced of item 5 because it did not fit. The analysis result using Cronbach's alpha shows that the overall reliability is very good, and the item reliability is good. Rating scale analysis using partial credit ratings and probability curves shows that respondents need help understanding the five-point Likert scale. The analysis of the item difficulty based on Logit and Wright maps shows that the most difficult item to work on is item 14. Items with moderate categories are items 13, 12, 1, 4, 10, 6, 2, 5, and 7. Items easy to work on are items 11, 15, and 9. The bias results show item 14 gender-biased. The results of the interaction between the item and the person through the ICC plot image show that all items are on the curve of the outfit confidence space and follow the Rasch modeling.

To cite this article (in APA style):
Prasetya, W., & Pratama, A. (2023). Item quality analysis using the Rasch model to measure critical thinking ability in the material of the human digestive system of Biology subject in high school. *Jurnal Penelitian dan Evaluasi Pendidikan, 27*(1), 76-91. doi:https://doi.org/10.21831/pep.v27i1.58873

INTRODUCTION

Science learning is expected to be able to guide students to develop 21st-century skills. One of the skills needed in the 21st century is critical thinking skills. They are important for student development and valuable for living in society (Danczak et al., 2017). The development of critical thinking skills needs to be continuously evaluated as a measure of the success of the learning process and as a reflective study material to improve the quality of learning.

The success of the teaching program can be seen from the results of the learning evaluation. There is one important component in the evaluation, namely, the test. A test is a tool for measuring the achievement of learning objectives (Widoyoko, 2009). The test as a measuring tool must be valid and reliable. A measuring instrument is said to be valid if it measures what it is supposed to measure without any significant bias or distortion (Matondang, 2009). Validity and reliability are affected by the instrument and the subject being measured. Invalid and reliable tests will give biased results and can even harm students (Widyaningsih & Yusuf, 2018).

One of the goals of learning biology is to apply what is learned to everyday life so students need to be trained to think critically to apply biology concepts in dealing with problems in everyday life. Critical thinking involves reasoning and the ability to separate biology and opinion (Chukwuyenum, 2013). A critical person will examine information before accepting or rejecting

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v27i1.58873

a solution to an existing problem. Practicing critical thinking skills becomes very important because it can help understand the logical relationship between ideas, build and evaluate arguments, and solve problems systematically (Riyanti et al., 2016).

The consequence of thinking that critical thinking skills are important in learning is that the teacher must provide an element of stimulation by creating an evaluation system that can open mindsets from remembering facts to critical thinking. Following its characteristics, critical thinking requires practice, one of which is working on evaluation questions that develop critical thinking skills (Kartimi, 2012).

Empirical data on the critical thinking skills of high school students show that there is a need for improvement. Studies show that students' critical thinking skills are generally in the low to moderate category, with some indicators being weaker than other indicators. For example, in a study, the evaluation, analysis, and self-regulation sub-skills were found to be the lowest critical thinking sub-skills mastered by students compared to other critical thinking sub-skills (Basri et al., 2019). However, some studies have found that certain learning models, such as the 5E Learning Cycle (Miarti et al., 2021) and the Conceptual Understanding Procedures (CUPs) learning model (Ariesta et al., 2019), can have a positive effect on students' critical thinking skills. In addition, research shows that the intellectual level (IQ) of students has a correlation with their critical thinking abilities (Hasanah et al., 2020). Overall, further research is needed to improve high school students' critical thinking skills.

Program for International Student Assessment (Center for Educational Assessment, 2018) proves that the science achievement score of Indonesian students is 396 out of 489. It shows that the average critical thinking ability of students in Indonesia is still relatively low. The PISA results obtained can be used to see the level of students' critical thinking because the questions contained in PISA are higher-order thinking skill (HOTS) questions; of course, this can encourage students to have the ability to solve problems, think critically, think creatively, reason, and make decision.

The results of interviews with biology teachers at a Yogyakarta high school found that critical thinking skills had never been measured using tests in the form of essay questions. However, educators stated that the average critical thinking was still needed to be improved because very few students actively asked questions while learning. Most students scored less than the minimum mastery criterion (MMC). For cognitive ability, the average did not meet expectations. For example, during midterm exams, the teacher expected that the students who could achieve MMC was 70%, but it turned out that only 50% could.

Observations on the discovery of questions used in one of the senior high schools in Yogyakarta showed that most teachers took sample questions from the handbook and the questions' validity and reliability were not measured. Thus, the quality of the questions used is unknown. Some of the questions given are low-order thinking skill questions, namely C1-C3, so students only memorize the material provided during the learning process and cannot understand the material properly. This allows students to be lazy to think. The teacher has not considered aspects of critical thinking skills because there is no test for critical thinking skills in biology teaching at school.

The importance of developing the HOTS instrument can be seen in efforts to measure and develop higher-order thinking skills which are urgently needed in today's education. HOTS instruments go beyond basic cognitive understanding and encourage students to apply, analyze, evaluate, and create knowledge in greater depth. By using the HOTS instrument, schools can evaluate students' abilities to explore more complex conceptual understandings, think critically, solve problems, and make decisions based on critical thinking. The development of HOTS instruments also helps provide teachers with valuable feedback to design more challenging and relevant lessons, and to prepare students to face challenges in the real world outside of school. HOTS instruments are important in developing students' skills needed to adapt, innovate, and contribute to an increasingly complex and rapidly changing society.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v27i1.58873

The use of the Rasch model to measure critical thinking skills has received attention in recent years. Several studies have used the Rasch model to analyze the quality of critical thinking items and categorize students' abilities. For example, using the Rasch model to measure the critical thinking skills of elementary school students in STEM learning (Hamdu et al., 2020), while finding that the Rasch model test shows that the experimental class has the ability to synthesize attitudes toward scientific inquiry (Wahyudiati, 2022). The Rasch model has also been used to validate critical thinking test items based on Buddhist philosophy (BCTA) (Susongko et al., 2022) and developing science learning handouts on the human digestive system to improve critical thinking skills (Sulastri et al., 2022). In addition, the Rasch model has been used to test critical thinking skills in in ecosystem materials (Karoror & Jalmo, 2022) and developing test instruments based on critical thinking skills that are integrated with Javanese cultural traditions in an Islamic context (Agustina et al., 2023). Overall, the Rasch model proves to be a useful tool in analyzing critical thinking skills and improving learning outcomes in various fields. Thus, the importance of measuring critical thinking skills and analyzing instruments using the Rasch model to assess critical thinking skills has become very important.

Accordingly, this study analyzed the instrument reliability using the Rasch model (Nielsen, 2018; Sumintono, 2018). The Rasch model is a statistical approach used to measure performance, perceptions, and attitudes (Bonsaksen et al., 2013; Nielsen, 2018). Evaluation of critical thinking skills with the Rasch model has more advantages than classical test theory because it can improve evaluation quality in quantitative and qualitative studies (Chan et al., 2014). Some of the advantages of using the Rasch model are: (1) it produces a linear and one-dimensional scale, (2) there is a need for conformity between the data and the measurement model, (3) it can call it calculates the standard error, (4) it can estimate the size of people and the level of difficulty of items through a linear scale that is similar to a standard unit (log), and (5) it can check the evaluation system logically and consistently (Planinic et al., 2019).

The Rasch model can analyze evaluation instruments based on several parameters. For the advantages of the Rasch model, an instrument must be tested for reliability, validity, differentiability, suitability, and level of difficulty using the Rasch model (Nielsen, 2018; Sumintono & Widhiarso, 2015). These stages are very important to obtain a reliable evaluation instrument. Therefore, it is necessary to analyze the instrument based on the needs in evaluating critical thinking skills, where the instrument reliability is analyzed using the Rasch model. Thus, this study aims to analyze the validity, reliability, scale comprehension, item difficulty, item bias, and interactions between items and persons through ICC plot images on creative thinking ability assessment instruments, and digestive system material using the Rasch model.

RESEARCH METHOD

This research was conducted as one of the stages of research and development of a problem-based learning virtual laboratory using the ADDIE model. The sampling technique used in this study is the simple random sampling technique. The participants of this study are 63 students of high schools in Yogyakarta. Characteristics of class and gender respondents are presented in Table 1. Data collection used a test that had been prepared according to the needs of critical thinking skills.

Table 1. Characteristic of Respondents

No.	Characteristic of Respondents	Category	Code	Number of Respondents
1.	Class	MIPA 1	1	32
		MIPA 2	2	31
2.	Gender	Male	1	24
		Female	2	39

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v27i1.58873

The test items are 15 essay questions on the material of the human digestive system using a five-point Likert scale. The indicators of critical thinking skills used include interpreting, analyzing, evaluating, explaining, and concluding (Facione, 1992). The collected data were analyzed using the Rasch model with the Winstep 5.0.3.4 program (Faradillah & Adlina, 2021). The research was conducted at a Yogyakarta high school in October 2022 using the Google Form application.

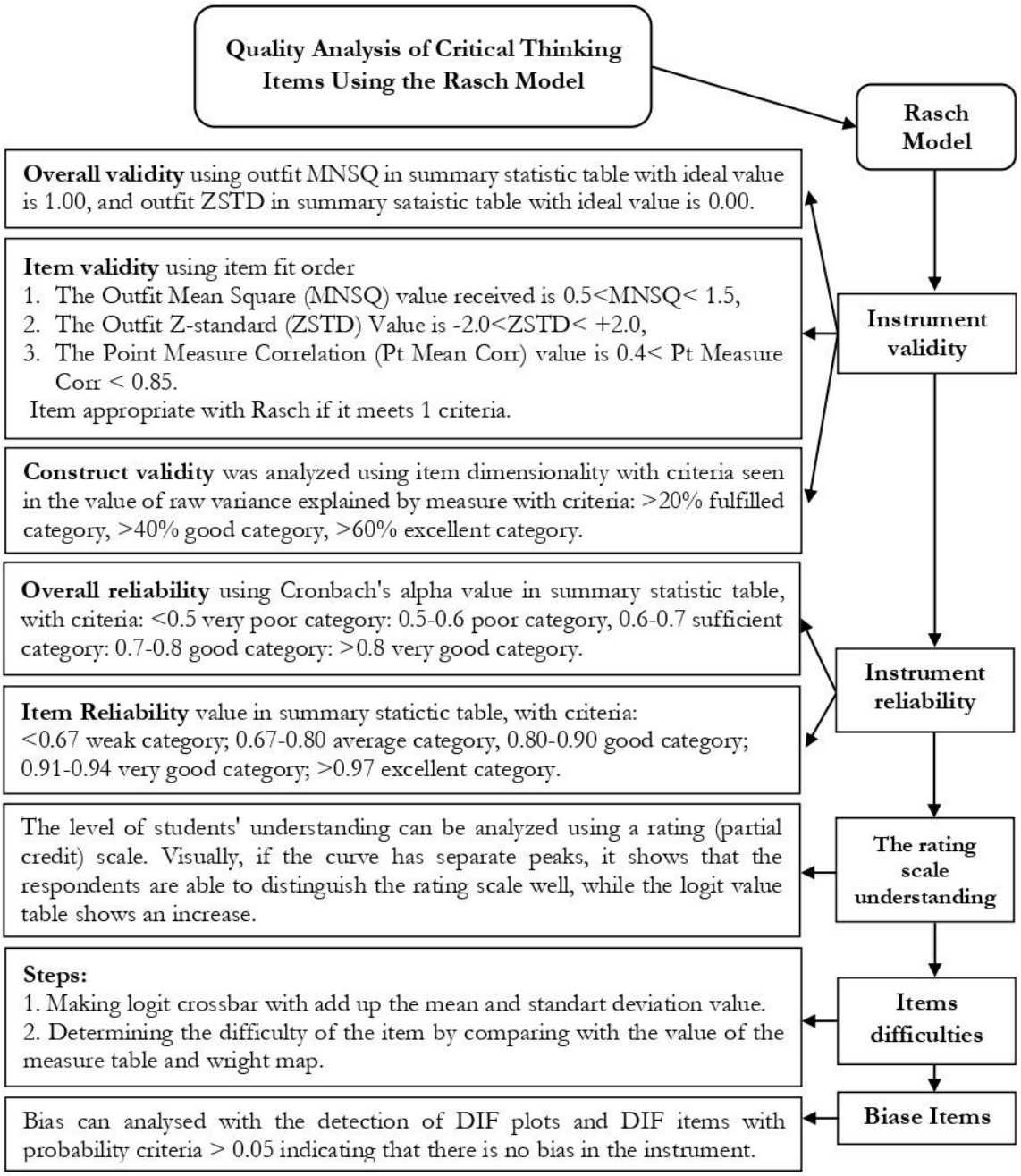

Figure 1. Instrument Analysis Process and Criteria Using Rasch Model (Austvoll-Dahlgren et al., 2017; Pontoppidan et al., 2018; Sumintono & Widhiarso, 2015)

The analysis phase begins with validity testing. Validity testing includes overall validity using summary statistics, item validity using item: fit order, and construct validity. Instrument reliability analysis was reviewed using alpha value and item reliability in the statistical summary test. Respondents understood the scale using the partial credit rating scale and probability curve.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v27i1.58873

Items are analyzed by creating a log bar to classify the item difficulty level based on the logit and wright map. Bias items were analyzed using DIF tables and plots. The interaction between items and persons is analyzed through ICC plots of images which have been modeled using probabilities developed by the Rasch model analysis, and are governed by two main parameters, namely the difficulty of the item and the ability of the person.

FINDINGS AND DISCUSSION

Instrument Validity Analysis

The validity test result is divided into two, namely, the overall validity of the instrument and the item (Planinic et al., 2019). The result of the analysis is presented in Table 2. The result of the analysis of the instrument's validity in Table 3 from the statistical summary shows whether the instrument is valid to use (Runco & Acar, 2012). Based on the value of the MNSQ outfit item (statement item), the instrument is suitable for evaluation because the result shows 1.00 and is included in the ideal value of 1.00. Based on the ZSTD item and person outfit values, the instrument shows that the data have a logical estimate because the result shows -0.28, close to the ideal value of 0.00 (Sumintono & Widhiarso, 2015).

Table 2. Summary Statistics of the Overall Validity Result (Kim, 2021; Sumintono, 2018)

No.	Value on Measurement	Measurement Type	Score	Value category
1.	Outfit MNSQ	Item	1.00	Accepted
2.	Outfit ZSTD	Item	-0.28	Accepted

Based on the American Educational Research Association (AERA) and American Psychological Association (APA), strong validity has evidence, and response validity is the reliability of the instrument when the respondent gives a response. The instrument's validity has used expert judgment, then directly used for the test. The validity test in the Rasch model informs the quality of the instrument so that the validity test is now more reliable (AERA & APA, 2014). The results of the item dimension test can be seen in Table 3, which shows that the instrument's construct validity has met the criteria.

Table 3. The Results of the Item Dimensionality to Analyze Construct Validity

Variance Explained by Measure	Value Category	Unexpected Variance 1st Contrast of PCA Residuals		Value Category
		Eigenvalue	Observed	
35.6%	Fulfilled	2.6	11.4%	Accepted

The results on the unexpected variance of 1 PCA residue contrast indicate that the criteria are accepted and that all statement items show conformity. The result is unidimensional. This means that the instrument can measure the range of variables or responses to questions to measure critical thinking skills. The construct validity of the instrument content variable can already measure what you want to know. Using the Rasch application model can determine the construct validity of the instrument. Based on research conducted by Madyani et al. (2020), construct validity has not been analyzed so this test has novelty. The results of the instrument reliability test can be seen from the suitability analysis of the items used to find out which items are not appropriate. Suitability analysis uses items that fit the order in Table 4. The results show that all items can be used to measure responses (Austvoll-Dahlgren et al., 2017). Based on Table 4, all items do not require revision to meet these criteria except for item 5, which must be repaired or replaced because it does not fit. The Rasch analysis model can direct instrument makers to revise the items that are not appropriate so that they have reliability in measurement.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v27i1.58873

Table 4. Recapitulation of Item Validity Values in Terms of Three Criteria

Item Number	Outfit MNSQ		OUTFIT ZSTD		PTMEA CORR	
	Mark	Information	Mark	Information	Mark	Information
14	1.08	Accepted	.37	Accepted	.35	Good
3	.74	Accepted	-1.24	Accepted	.52	Very good
8	1.21	Accepted	.97	Accepted	.25	Fair
13	.57	Accepted	-2.48	Bad	.67	Very good
12	.86	Accepted	-.67	Accepted	.57	Very good
1	1.00	Accepted	.07	Accepted	.55	Very good
4	1.20	Accepted	1.03	Accepted	.58	Very good
10	1.06	Accepted	.36	Accepted	.44	Very good
6	.93	Accepted	-.30	Bad	.47	Very good
2	.67	Accepted	-2.14	Bad	.65	Very good
5	2.09	Bad	5.01	Bad	.22	Fair
7	.36	Accepted	-5.06	Bad	.71	Very good
11	.80	Accepted	-1.20	Accepted	.74	Very good
15	1.80	Accepted	3.96	Bad	.37	Good
9	.59	Accepted	-2.83	Bad	.79	Very good

Instrument Reliability Analysis

Table 5 shows the results of the reliability test. The overall instrument with a Cronbach alpha value of 0.81 has a very good category. The reliability of the item is 0.88 which has a very good category (Sumintono & Widhiarso, 2015). The instrument has consistent results when tested on the population (Plucker et al., 2014; Runco & Albert, 1985). The grouping of items has a very good category because there are three categories of item difficulty levels on the instrument. The grouping of respondents is in a good category because the respondents have five ability levels. Thus, the instrument can be used to determine the grouping of the items and respondents in evaluating critical thinking skills (Göçmen & Coşkun, 2019; Sumintono & Widhiarso, 2015).

Table 5. The Results of the Summary Statistics Measurement to Analyze the Reliability Value and Grouping of the Items and Respondents

No.	Value on Measurement	Measurement Type	Score	Value category
1.	Alpha	Overall	0.81	Very good
2.	Reliability	Items	0.88	Good

The Rating Scale Understanding Analysis

Rating scale evaluation (1, 2, 3, 4, 5) can be seen from the peaks of each scale on the probability curve in Figure 2, which shows separate peaks; details can be seen in Figure 3. Table 6 shows that the scale five ratings (excellent, good, sufficient, low, and very low) on the instrument did not work well if respondents understood each scale category in the questions given, meaning that the rating scale did not work well. One of the characteristics that the rating scale has not functioned properly is looking at it based on the Andrich Threshold which does not show a monotonous increase (Sumintono & Widhiarso, 2015).

Figure 2 represents the probability curve of the rating scale instrument. This curve depicts the probability distribution for each rating scale on the instrument for analyzing critical thinking skills in the human digestive system. In this context, the probability curve provides information about how often or how likely a particular rating scale is used by respondents when providing assessments or ratings for the given questions or items. In the analysis of this probability curve, each rating scale (1, 2, 3, 4, 5) has distinct peaks. This indicates that respondents have clear tendencies in using specific rating scales when providing answers or assessments in the instrument. Through this probability curve, we can gain insights into the preferences or tendencies of re-

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v27i1.58873

spondents to use the provided rating scale (Hansen & Kjaersgaard, 2020; Lin et al., 2017). This can be helpful in evaluating the quality of the instrument and understanding how respondents respond to the questions or items in the instrument.

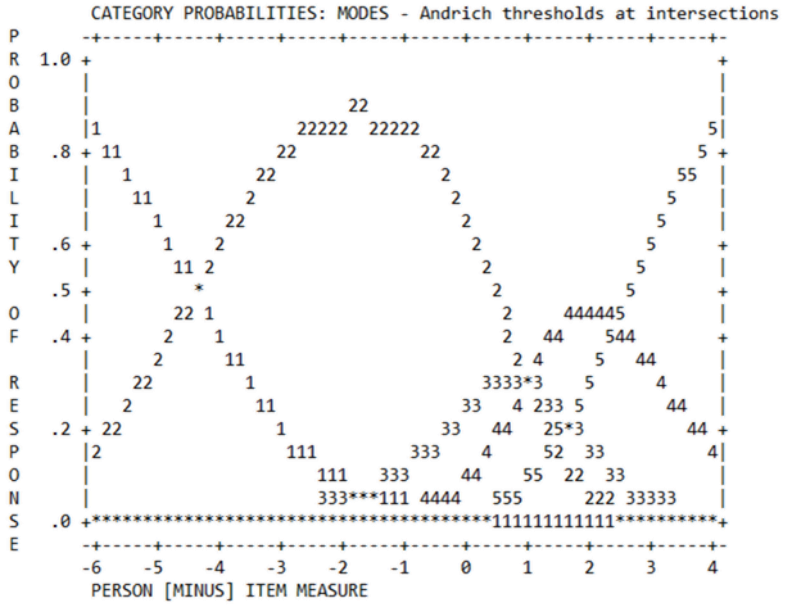

Figure 2. Probability Curve of the Rating Scale Instrument

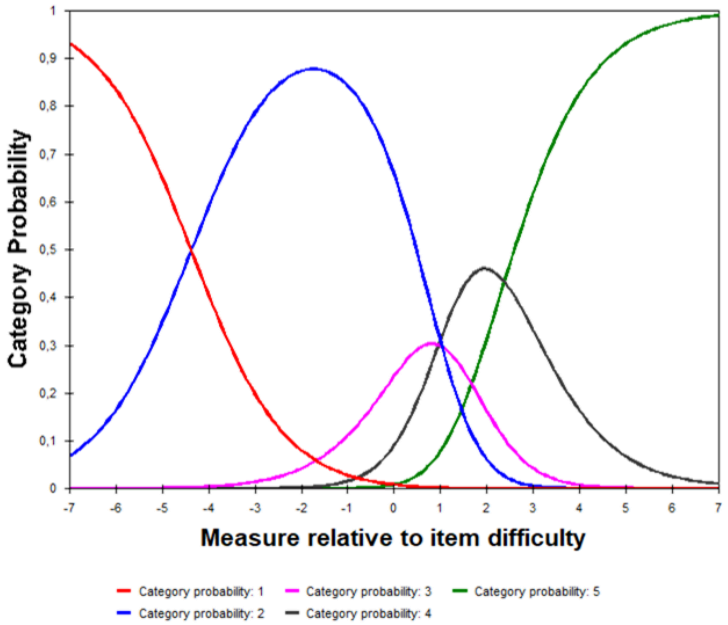

Figure 3. Probability Curve of the Rating Scale Instrument

Figure 3 presents probability curves where the y-axis represents the probability of each category of the five response options at each level of item difficulty on the x-axis. Each color corresponds to a different response option: 1 = red, 2 = blue, 3 = pink, 4 = black, 5 = green. The intersections of adjacent curves are the thresholds. The category probability curves for the item above have disordered thresholds. The responses to this item are not distributed in a logically progressive order, and categories 3 and 4 do not have a high probability.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v27i1.58873

Table 6. The Statistics of Rating Scale Analysis

Category Label	Observed		SE	Andrich Threshold	Category Measure
	Count	Frequency %			
1 (Very Low)	6	1	-	None	-5.48
2 (Low)	392	41	0.41	-4.37	-1.74
3 (Sufficient)	226	24	0.08	+1.03	+0.83
4 (Good)	232	25	0.08	+0.96	+1.96
5 (Excellent)	89	9	0.12	+2.38	+3.64

Understanding of the scale can be seen through statistics on the Rasch model. The scale is only analyzed through descriptive analysis when the respondent fills in all the questions. The researcher concluded that the respondents could not understand the rating scale properly. So far, the understanding of the scale has not been tested. A bad probability curve can be used to analyze a lack of scale, for example, by reducing the scale range or eliminating meaningful neutral ratings.

Rasch analysis also provides information about the number of best response categories on the scale. To analyze whether the category calibration increases regularly, the response options are assessed by the category probability curve which can be seen in Figure 3 (Linacre, 2002). They indicate the likelihood that a subject with a certain person measure relative to item difficulty will choose a category (Pesudovs et al., 2007). The threshold is the midpoint between adjacent response categories thereby expressing the point at which the probability of selecting either response category is equal (McAlinden et al., 2012). If disordered thresholds occur, the situation must be changed by collapsing the required categories into adjacent categories (Andrich, 2013b, 2013a; Pesudovs et al., 2007). To detect this situation, the Andrich threshold measurements must be checked so that the thresholds must be spaced at least 1.4 logits (Linacre, 2002). Item reduction is made iterative so that one item is removed at a time (Pesudovs et al., 2003). Thus, when an item is omitted, the fit to the model is consequently re-estimated because it has been shown that the fit is relative so omitting an item leads to variation in the fit. Then, items with the most candidate criteria, sorted by priority, are eliminated first (Cantó-Cerdán et al., 2021).

The Analysis of the Difficulty Level of the Items

Assessment of the difficulty level of the items is done by using item size. Separation or difficulty level of the items is determined by adding up the average value with the standard deviation (0.00+0.45= 0.45 which is used to create a log bar. The log bar can be used to determine difficult, moderate, and easy items and outliers, as seen in Table 7. It is necessary to pay attention to the values of item and person measurements to determine the response given. The results of the person measurement show an average value (M) of 0.02 logit on the measurement item, namely -0.28 logit. Thus, the ability to answer is above the average difficulty level of standard statement items (Sumintono & Widhiarso, 2015).

Logit bar in Table 7, then integrated into the measure items in Table 7 and the wright map in Figure 4, to determine the item question's classification. The results showed that the difficulty level of the items indicated the order of the items from the most difficult items to be worked on by respondents to the easiest items to work on. The table that item 14 is the most difficult to work on or has the highest level of difficulty because it has the highest level of difficulty, so it is included in the outlier item category. Then the items with a medium category are items 13, 12, 1, 4, 10, 6, 2, 5, and 7. The items that are easy to work on or items with a low level of difficulty are items 11, 15, and 9. The difficulty level of items can be categorized by adding up the mean value and standard deviation. In the table, the sum of the mean and standard deviation values is equal to 0.45. Items that exceed the logit value of 0.90 or less than -0.90 indicate the item is an outlier or cannot be used and needs to be discarded.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v27i1.58873

Table 7. Grouping of Item Difficulty Levels

Item Group	Category	Item
Outliers	logit > 0.90	14
Difficult	(0.90 < logit < 0.45)	3, 8
Moderate	(0.45 ≤ logit ≤ -0.45)	13, 12, 1, 4, 10, 6, 2, 5, 7
Easy	(-0.90 < logit < -0.45)	11, 15, 9
Outliers	logit < -0.90	0

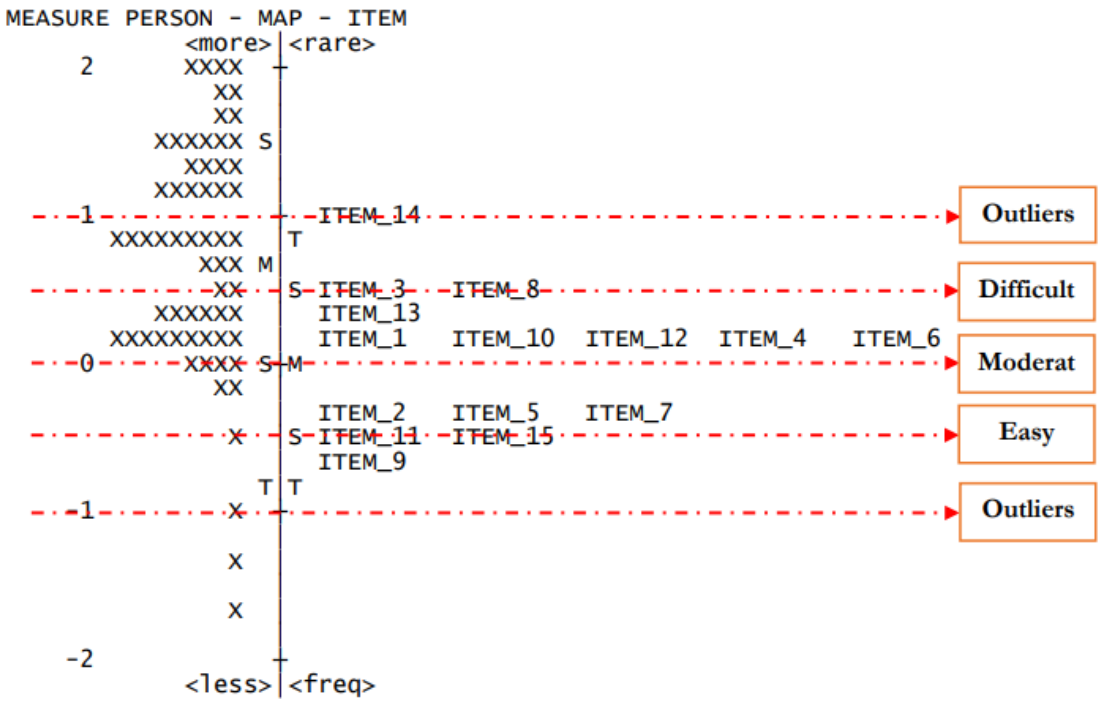

Figure 4. Wright Item Map to Analyze the Difficulty Level of the Questions

The Analysis of Gender-Bias Items

The result of the bias test using DIF in Figure 5 shows a probability with a bias criterion of <0.05. The blue L code indicates the male gender, and the red P code indicates the female gender. One item was found to be seen from the gender factor, namely item 14. DIF analysis was used to confirm whether there were items that had a gender bias (women and men) that affected critical thinking skills on the material of the human digestive system. DIF analysis can identify participant bias based on subgroups or variables for each item in the instrument used (Boone et al., 2014; Khine, 2020). DIF was determined following two categories: significant probability ($p < 0.05$) and DIF contrast. Three DIF contrast classifications (Zwick et al., 1999) are negligible, slight to moderate (|DIF | 0.43 log), and moderate to large (|DIF | 0.64 log). Figure 5 shows that having items that fall into the DIF category is based on a significant probability. Thus, we can conclude that item 14 is categorized as DIF indicating that the instrument has a bias problem.

The farthest distance of the items from the average indicates a significant difference in difficulty level between men and women (Azizah et al., 2022). In this case, there are groups that benefit more and there are groups that are disadvantaged because a problem seems more difficult for women than for men. The graph shows that item 14 is more favorable for males than females. When these items were confirmed, it turned out that the narratives on the questions that had been prepared gave different assumptions to the respondents and had an impact on gender status. This is in line with several studies that have used Rasch analysis to investigate

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v27i1.58873

gender bias in various scales and questionnaires. This study has found evidence of gender bias in items using Rasch analysis. For example, finding the uniform differential item function (DIF) for two items in the Patient Assessed Elbow Evaluation questionnaire, with one item indicating the DIF for gender (Vincent et al., 2015). Similarly, we found that several items in the STEM Career Interest Survey detect gender bias (Ardianto et al., 2023). It was found that the three items on the GAIN substance problem scale appeared different between men and women (Claro et al., 2015). For this reason, it is very important to compose narratives and choose words so that the resulting items or questions do not lead to different assumptions about gender, so as not to cause gender bias.

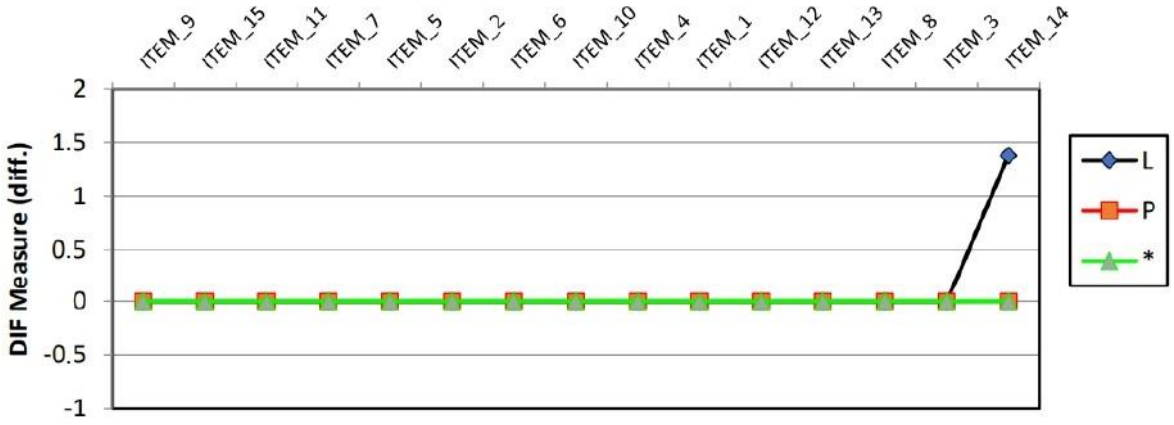

Figure 5. DIF Plot of Gender Bias on the Significance of Answering Items

Investigating the Interaction between Item and Person via ICC Plot

Item and person interaction analysis on the Rasch plot ICC model is an important aspect in evaluating the psychometric properties of measurement instruments. Research by Planinic et al. (2019) which uses the Rasch model reveals that it is a probabilistic model that describes the interaction of a person with a test or survey item and is governed by two parameters: item difficulty and person's ability. Figure 6 shows the ICC plot for critical thinking which explains that the red line indicates the ideal line of the Rasch model or the line modeled with the probability developed by Rasch analysis. The red line illustrates that the higher it gets indicates the item's difficulty level. On the other hand, the blue line shows the data or distribution of answers from respondents. Overall, the person is considered capable of answering all items. This can also be confirmed by the results presented in Figure 4. Wright Item Map, although there are items that are not fit and must be replaced or discarded, overall, based on testing using the ICC expected score, shows that there are no items that show an inappropriate response. In other words, all items are on the curve of the outfit confidence space. This means that the 15 items in the critical thinking ability instrument on the material of the human digestive system are able to measure exactly what is being measured.

The research conducted by McCamey (2014) revealed the Rasch model allows estimation of item difficulty and person's ability on a general scale, which is useful for evaluating the reliability and validity of measurement instruments. The person-item distribution map (PIDM) is a graphical representation of the interactions between persons and items, which can provide meaningful information about the effectiveness of student learning (Nopiah et al., 2012). The Rasch model is also useful for examining individuals in terms of how they respond to items using a person map (Subroto et al., 2022). The Rasch model is designed to find the characteristics of persons and items that are independent of each other and interval levels, and it can generalize the characteristics of persons and items outside of a sample of respondents and certain items (Benson et al., 2018).

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v27i1.58873

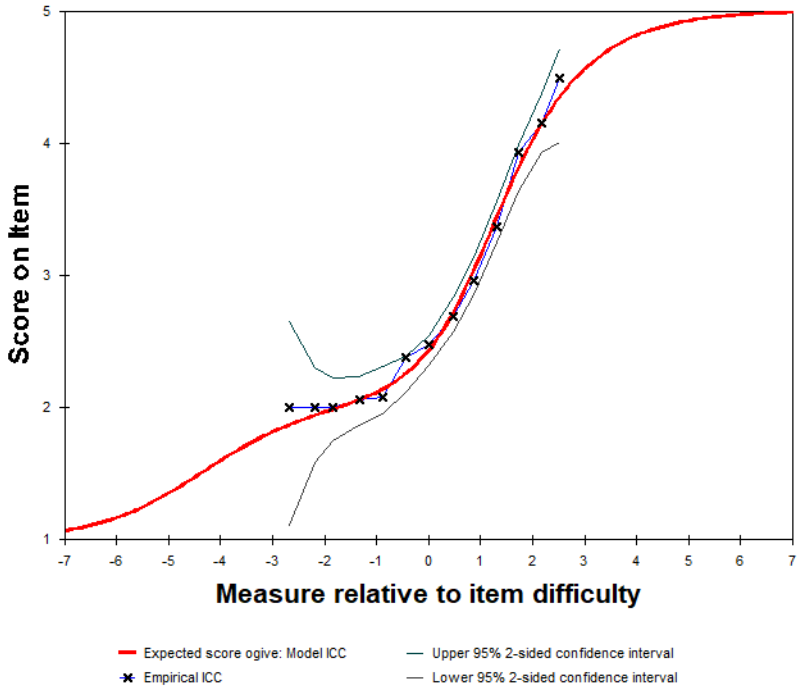

Figure 6. Interaction between Item and Person through ICC Plot

On the other hand, in the study by Rifbjerg-Madsen et al. (2017) Rasch analysis revealed acceptable psychometric rating scale properties, including calibration of thresholds, no interference in category size for all items, and small distances between thresholds for all items. Likewise, in the study of Imani et al. (2018) Rasch analysis shows that the Epworth Sleepiness Scale for Children and Adolescents is a reliable and valid instrument. Internal consistency, test-retest reliability, and Rasch analysis show that the instrument is reliable and valid. Therefore, item and person interaction analysis on the Rasch plot ICC model is a valuable tool for evaluating the psychometric properties of measurement instruments.

CONCLUSION

Based on the result of this study, it can be concluded that the instrument for measuring critical thinking skills in the material of the human digestive system has good criteria when applied in the teaching-learning process. Based on the analysis, the overall validity is acceptable, and the item validity requires improvement; item 5 must be repaired or replaced because the item does not fit. Based on the results of the analysis, the overall reliability is very good, and the item reliability is good. Rating scale analysis shows that respondents do not understand the five-point Linkert scale, so a bad probability curve can be used to analyze the lack of a scale, for example, by reducing the scale range or eliminating a meaningful neutral rating. The analysis of the item difficulty showed that the item 14 was the most difficult to work on or had the highest level of difficulty because it had the highest level of difficulty so that it was included in the outlier item category. The items with a moderate category are items 13, 12, 1, 4, 10, 6, 2, 5, and 7. The items that are easy to do or items with a low level of difficulty are items 11, 15, and 9. The bias results show that one item is found in terms of the gender factor, namely item 14. Moreover, the results of the interaction between the item and the person through the ICC plot image show that even though there are items that do not fit and must be replaced or discarded, overall it is based on testing using the expected score ICC, showing that there are no items that show inappropriate responses. In other words, all items are on the curve of the outfit confidence space and are by the Rasch modeling.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v27i1.58873

REFERENCES

AERA & APA. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Agustina, D. F., Raharjo, R., Isnawati, I., & Hartono, D. (2023). Test instrument based on critical thinking skills integrated Javanese cultural tradition in Islamic context. *International Journal of Social Science And Human Research,* 6(2), 987-995. https://doi.org/10.47191/ijsshr/v6-i2-30

Andrich, D. (2013a). An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any "Threshold Disorder Controversy." *Educational and Psychological Measurement*, *73*(1), 78–124. https://doi.org/10.1177/0013164412450877

Andrich, D. (2013b). The legacies of R. A. Fisher and K. Pearson in the application of the polytomous Rasch model for assessing the empirical ordering of categories. *Educational and Psychological Measurement*, *73*(4), 553–580. https://doi.org/10.1177/0013164413477107

Ardianto, D., Rubini, B., & Pursitasari, I. D. (2023). Assessing STEM career interest among secondary students: A Rasch model measurement analysis. *Eurasia Journal of Mathematics, Science and Technology Education*, *19*(1), em2213. https://doi.org/10.29333/ejmste/12796

Ariesta, P., Susanti, R., & Rahayu, E. S. (2019). The influence of Conceptual Understanding Procedures (CuPs) learning model with (the use of) Bio-Quartet cards. *J.Biol.Educ., 8*(1), 50–55. http://journal.unnes.ac.id/sju/index.php/ujbe

Austvoll-Dahlgren, A., Guttersrud, Ø., Nsangi, A., Semakula, D., & Oxman, A. D. (2017). Measuring ability to assess claims about treatment effects: A latent trait analysis of items from the "Claim Evaluation Tools" database using Rasch modelling. *BMJ Open*, *7*(5), e013185. https://doi.org/10.1136/bmjopen-2016-013185

Azizah, N., Suseno, M., & Hayat, B. (2022). Item analysis of the Rasch model items in the final semester exam indonesian language lesson. *World Journal of English Language*, *12*(1), 15–26. https://doi.org/10.5430/wjel.v12n1p15

Basri, H., Purwanto, P., As'ari, A. R., & Sisworo, S. (2019). Investigating critical thinking skill of junior high school in solving mathematical problem. *International Journal of Instruction*, *12*(3), 745–758. https://doi.org/10.29333/iji.2019.12345a

Benson, N. F., Beaujean, A. A., Donohue, A., & Ward, E. (2018). W Scores: Background and derivation. *Journal of Psychoeducational Assessment*, *36*(3), 273–277. https://doi.org/10.1177/0734282916677433

Bonsaksen, T., Kottorp, A., Gay, C., Fagermoen, M. S., & Lerdal, A. (2013). Rasch analysis of the general self-efficacy scale in a sample of persons with morbid obesity. *Health and Quality of Life Outcomes, 11*, 202 https://doi.org/10.1186/1477-7525-11-202

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

Cantó-Cerdán, M., Cacho-Martínez, P., Lara-Lacárcel, F., & García-Muñoz, Á. (2021). Rasch analysis for development and reduction of Symptom Questionnaire for Visual Dysfunctions (SQVD). *Scientific Reports*, *11*(1), 14855. https://doi.org/10.1038/s41598-021-94166-9

Center for Educational Assessment. (2018). *Pendidikan di Indonesia: Belajar dari hasil PISA 2018 programme for international student assessment*. Center for Educational Assessment, Badan Research and Development Agency, Ministry of Education and Culture.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v27i1.58873

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. https://doi.org/10.1016/j.sbspro.2014.03.658

Chukwuyenum, A. N. (2013). Impact of critical thinking on performance in Mathematics among senior secondary school students in Lagos State. *IOSR Journal of Research & Method in Education, 3*(5), 27910355. https://doi.org/10.9790/7388-0351825

Danczak, S. M., Thompson, C. D., & Overton, T. L. (2017). What does the term critical thinking mean to you? A qualitative analysis of chemistry undergraduate, teaching staff and employers' views of critical thinking. *Chemistry Education Research and Practice*, *18*(3), 420–434. https://doi.org/10.1039/c6rp00249h

Facione, P. A. (1992). *Critical thinking: What it is and why it counts*. Insight Assessment.

Faradillah, A., & Adlina, S. (2021). Validity of critical thinking skills instrument on prospective Mathematics teachers. *Jurnal Penelitian dan Evaluasi Pendidikan*, *25*(2), 126-137. https://doi.org/10.21831/pep.v25i2.40662

Claro, H. C., de Oliveira, M. A. F., Fernandes, I. F. A. L., Titus, J. C., Tarifa, R. R., Rojas, T. F., & Pinho, P. H. (2015). Rasch model of the GAIN substance problem scale among inpatient and outpatient clients in the city of São Paulo, Brazil. *Addictive Behaviors Reports*, *2*, 55–60. https://doi.org/10.1016/j.abrep.2015.08.001

Göçmen, Ö., & Coşkun, H. (2019). The effects of the six thinking hats and speed on creativity in brainstorming. *Thinking Skills and Creativity*, *31*, 284–295. https://doi.org/10.1016/j.tsc.2019.02.006

Hamdu, G., Fuadi, F. N., Yulianto, A., & Akhirani, Y. S. (2020). Items quality analysis using Rasch model to measure elementary school students' critical thinking skill on Stem learning. *JPI (Jurnal Pendidikan Indonesia)*, *9*(1), 61-74. https://doi.org/10.23887/jpi-undiksha.v9i1.20884

Hansen, T., & Kjaersgaard, A. (2020). Item analysis of the Eating Assessment Tool (EAT-10) by the Rasch model: A secondary analysis of cross-sectional survey data obtained among community-dwelling elders. *Health and Quality of Life Outcomes, 18*(1), 1–14. https://doi.org/10.1186/s12955-020-01384-2

Hasanah, S. N., Sunarno, W., & Prayitno, B. A. (2020). Profile of students' critical thinking skills in junior high schools in Surakarta. In *Proceedings of the 3rd International Conference on Learning Innovation and Quality Education (ICLIQE 2019),* pp. 570-575. https://doi.org/10.2991/assehr.k.200129.070

Imani, V., Lin, C. Y., Jalilolghadr, S., & Pakpour, A. H. (2018). Factor structure and psychometric properties of a Persian translation of the Epworth Sleepiness Scale for children and adolescents. *Health Promotion Perspectives*, *8*(3), 200–207. https://doi.org/10.15171/hpp.2018.27

Karoror, I., & Jalmo, T. (2022). Profile of critical thinking ability in Ecosystem materials using the Rasch model. *Jurnal Penelitian Pendidikan IPA, 3*(8), 1599-1604. https://doi.org/10.29303/jppipa.v8i3.1394

Kartimi, K. (2012). Pengembangan alat ukur berpikir kritis pada konsep Termokimia untuk siswa SMA. *Jurnal Scientiae Educatia*, *1*(1), 1-14. https://www.syekhnurjati.ac.id/jurnal/index.php/sceducatia/article/view/501

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v27i1.58873

Khine, M. S. (2020). Rasch measurement: Applications in quantitative educational research. In *Rasch measurement: Applications in quantitative educational research*. Springer Singapore. https://doi.org/10.1007/978-981-15-1800-3

Kim, J. (2021). Development and validation of the career adaptability scale for undergraduates in Korea. *Sustainability (Switzerland)*, *13*(19), 11004. https://doi.org/10.3390/su131911004

Lin, C. Y., Broström, A., Nilsen, P., Griffiths, M. D., & Pakpour, A. H. (2017). Psychometric validation of the Persian bergen social media addiction scale using classic test theory and Rasch models. *Journal of Behavioral Addictions,* *6*(4), 620–629. https://doi.org/10.1556/2006.6.2017.071

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3(1), 85-106. https://europepmc.org/article/med/11997586

Madyani, I., Yamtinah, S., Utomo, S. B., Saputro, S., & Mahardiani, L. (2020). Profile of students' creative thinking skills in science learning. In *Proceedings of the 3rd International Conference on Learning Innovation and Quality Education (ICLIQE 2019),* pp. 957-964. https://doi.org/10.2991/assehr.k.200129.119

Matondang, Z. (2009). Validitas dan reliabilitas suatu instrumen penelitian. *Jurnal Tabularasa PPs Unimed,* *6*(1), 87-97. http://digilib.unimed.ac.id/705/1/Validitas%20dan%20reliabilitas%20suatu%20instrumen%20penelitian.pdf

McAlinden, C., Khadka, J., Santos Paranhos, J. de F., Schor, P., & Pesudovs, K. (2012). Psychometric properties of the NEI-RQL-42 questionnaire in keratoconus. *Investigative Ophthalmology and Visual Science*, *53*(11), 7370–7374. https://doi.org/10.1167/iovs.12-9969

McCamey, R. (2014). A primer on the one-parameter Rasch model. *American Journal of Economics and Business Administration*, *6*(4), 159–163. https://doi.org/10.3844/ajebasp.2014.159.163

Miarti, E., Hasnunidah, N., & Abdurrahman, A. (2021). The effect of learning cycle 5E on critical thinking skills for junior high school students. *Scientiae Educatia*, *10*(2), 177. https://doi.org/10.24235/sc.educatia.v10i2.9127

Nielsen, T. (2018). The intrinsic and extrinsic motivation subscales of the motivated strategies for learning questionnaire: A Rasch-based construct validity study. *Cogent Education*, *5*(1), 1504485. https://doi.org/10.1080/2331186X.2018.1504485

Nopiah, Z. M., Rosli, S., Baharin, M. N., Othman, H., & Ismail, A. (2012). Evaluation of pre-assessment method on improving student's performance in complex analysis course. *Asian Social Science*, *8*(16), 134–139. https://doi.org/10.5539/ass.v8n16p134

Pesudovs, K., Burr, J. M., Harley, C., & Elliott, D. B. (2007). The development, assessment, and selection of questionnaires. *Optometry and Vision Science*, *84*(8), 663–674. https://doi.org/10.1097/OPX.0b013e318141fe75

Pesudovs, K., Garamendi, E., Keeves, J. P., & Elliott, D. B. (2003). The activities of daily vision scale for cataract surgery outcomes: Re-evaluating validity with Rasch analysis. *Investigative Ophthalmology and Visual Science*, *44*(7), 2892–2899. https://doi.org/10.1167/iovs.02-1075

Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research*, *15*(2), 020111. https://doi.org/10.1103/PhysRevPhysEducRes.15.020111

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v27i1.58873

Plucker, J. A., Qian, M., & Schmalensee, S. L. (2014). Is what you see what you really get? Comparison of scoring techniques in the assessment of real-world divergent thinking. *Creativity Research Journal*, *26*(2), 135–143. https://doi.org/10.1080/10400419.2014.901023

Pontoppidan, M., Nielsen, T., & Kristensen, I. H. (2018). Psychometric properties of the Danish parental stress scale: Rasch analysis in a sample of mothers with infants. *PLoS ONE*, *13*(11), e0205662. https://doi.org/10.1371/journal.pone.0205662

Rifbjerg-Madsen, S., Wæhrens, E. E., Danneskiold-Samsøe, B., & Amris, K. (2017). Psychometric properties of the painDETECT questionnaire in rheumatoid arthritis, psoriatic arthritis and spondyloarthritis: Rasch analysis and test-retest reliability. *Health and Quality of Life Outcomes*, *15*(1), 110. https://doi.org/10.1186/s12955-017-0681-1

Riyanti, A., Widiyatmoko, A., & Wusqo, I. U. (2016). pengaruh model pembelajaran kooperatif tipe Team Assisted Individualization berbantuan peta konsep terhadap hasil belajar dan keterampilan berpikir kritis siswa SMP tema Kalor. *Unnes Science Education Journal*, *5*(2), 70805795–70850229. http://journal.unnes.ac.id/sju/index.php/usej

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity Research Journal*, 24(1), 66–75. https://doi.org/10.1080/10400419.2012.652929

Runco, M. A., & Albert, R. S. (1985). The reliability and validity of ideational originality in the divergent thinking of academically gifted and nongifted children. *Educational and Psychological Measurement*, *45*(3), 483–501. https://doi.org/10.1177/001316448504500306

Subroto, G., Agust, S., Angela, A., Dezar, A., Zahra, D., Mirarizka, D., Rianto, F., Rayani, V., & Candra, M. (2022). Coastal students' perspectives on digital reading comprehension: A Rasch model analysis. In *Proceedings of the 1st International Conference on Maritime Education, ICOME 2021, 3-5 November 2021, Tanjungpinang, Riau Islands, Indonesia*. https://doi.org/10.4108/eai.3-11-2021.2314832

Sulastri, A., Badruzsaufari, B., Dharmono, D., Aufa, M. N., & Saputra, M. A. (2022). Development of Science handouts based on critical thinking skills on the topic of the Human Digestive System. *Jurnal Penelitian Pendidikan IPA*, *8*(2), 475–480. https://doi.org/10.29303/jppipa.v8i2.1156

Sumintono, B. (2018). Rasch model measurements as tools in assessment for learning. In *Proceedings of the 1st International Conference on Education Innovation (ICEI 2017),* pp. 38-42. https://doi.org/10.2991/icei-17.2018.11

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan.* Trim Komunikata.

Susongko, P., Yuenyong, C., & Zainudin, A. (2022). Buddhist critical thinking assessment using Rasch model. *Kasetsart Journal of Social Sciences, 43*(2), 285–292. https://doi.org/10.34044/j.kjss.2022.43.2.04

Vincent, J. I., MacDermid, J. C., King, G. J. W., & Grewal, R. (2015). Rasch analysis of the Patient Rated Elbow Evaluation questionnaire. *Health and Quality of Life Outcomes*, *13*(1), 84. https://doi.org/10.1186/s12955-015-0275-8

Wahyudiati, D. (2022). Critical thinking skills and scientific attitudes of pre-service Chemistry teachers through the implementation of problem-based learning model. *Jurnal Penelitian Pendidikan IPA*, *8*(1), 216–221. https://doi.org/10.29303/jppipa.v8i1.1278

Widoyoko, E. P. (2009). *Evaluasi program pembelajaran*. Pustaka Pelajar.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v27i1.58873

Widyaningsih, W., & Yusuf, I. (2018). Project based learning model based on simple teaching tools and critical thinking skills. *Physics Education Journal*, *1*(1), 12–21. https://doi.org/10.37891/kpej.v1i1.33

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, *36*(1), 1-28. https://www.jstor.org/stable/1435320

ISSN (print) 2685-7111 | ISSN (online) 2338-6061