25 (4), 2024, 1893-1905

Jurnal Pendidikan MIPA

e-ISSN: 2685-5488 | p-ISSN: 1411-2531

http://jurnal.fkip.unila.ac.id/index.php/jpmipa/

Development of a Complex Thinking Test Instrument for Decision Making of Environmental Change Topics

Anisyah Yuniarti[1,*], Afandi[1], Asih Triyanti[1], & Wiwit Artika[2]

[1]Department of Biology Education, Tanjungpura University, Indonesia

[2]Department of Biology Education, Syiah Kuala University, Indonesia

Abstract: The development of test instruments to measure students' decision-making abilities is rarely used in learning. This research aims to develop a complex thinking test instrument for decision-making aspects of environmental change topics. This type of research is research and development (R&D) using a 4-D model modified to 3-D (definition, design, and development) and combined with test instrument development using six stages. Interviews, test instrument validation sheets, and tests were conducted during research data collection. The test instrument developed consisted of 12 descriptive questions and was tested on 363 high school students. Data analysis consisted of content validity, inter-rater reliability, and Rasch analysis. Before the test instrument was tested, it was validated by seven experts. The research results show that the content validity analysis developed is valid because of V-count =0.94 > (V-table =0.76). Interrater reliability analysis obtained a value of 0.806, categorized as good. Item validity (item fit) shows 12 valid questions. The difficulty level of the questions (item measure) shows that three questions are challenging, three are complicated, four are easy, and two are straightforward. Reliability shows an excellent Cronbach's alpha value (0.89), a reasonable person reliability value (0.85), and an excellent item reliability value (0.98). Based on the analysis results, it can be concluded that the questions are declared valid and reliable.

Keywords: development, test instrument, decision making, presseisen taxonomy, rasch analysis.

▪ INTRODUCTION

Test instrument development is an activity that involves creating and compiling measuring instruments that can be used to measure students' ability levels. Test instruments are tools used to collect data in research (Nasution, 2016) ;Elan et al., 2022)). With test instruments in research, data can be collected (Magdalena et al., 2021). Student learning outcomes are obtained from test instruments created to see their thinking abilities.

Thinking is the act of retrieving knowledge that has been stored in memory to organize and summarize information. The ability to think is one of the abilities that needs to be developed to face the challenges of the 21st century (Mufidah & Wijaya, 2017). Therefore, the ability to think complexly or at a high level is essential for every student, both at school and in everyday life. Decision-making skills is one of the complex thinking abilities that need to be given and developed to students. Decision-making skills are a person's ability to think about and choose the best option from the many available options by considering the benefits and risks (Presseisen, 1985). The ability to make decisions in the learning process is due to the shift in the 21st-century learning paradigm (Sanjaya et al., 2019).

Students' complex thinking abilities in decision-making can be trained by providing test instruments in the form of questions that encourage students to think complexly about their decision-making abilities. Designing good test instruments appropriate to the level of thinking ability can improve students' high-level thinking abilities (Afrita &

Anisyah Yuniarti
*Email: anisyah.yuniarti@fkip.untan.ac.id

DOI: http://dx.doi.org/10.23960/jpmipa/v25i4.pp1893-1905
Received: 09 January 2025
Accepted: 20 January 2025
Published: 25 January 2025

Rahmawati, 2020). Based on the results of interviews conducted by researchers with biology teachers at SMAN 3, SMAN 5, SMAN 9, and SMAN 11, it was found that no test instrument could measure complex thinking abilities, especially in the aspect of decision-making. In line with research conducted by Maulana & Rochintaniawati (2021), it is stated that the learning process used has not directed students to practice decision-making skills, so that students' decision-making skills are not well developed. Teachers' test instruments still focus on Bloom's taxonomy with cognitive levels C2 (understanding) and C3 (applying). It makes students familiar with low-level thinking questions.

The factor that causes students' thinking abilities to be still low is the lack of training of Indonesian children in completing tests or questions that are analytical, evaluation and creative (Akmala et al., 2019). The low thinking ability of students is also because many teachers still choose the lecture method as the primary way of teaching (Suratno et al., 2020). Students' ability to think in complex decision-making aspects is still relatively low, as demonstrated by several students at Sidoarjo High School regarding difficulties in decision-making (Rahmasari et al., 2023). A decision-focused complex thinking test instrument for high school students is essential as it measures critical and analytical thinking abilities, which support students' life skills in dealing with educational, career, and social life choices. In addition, it helps teachers customize their learning methods according to students' needs and promotes character development by considering values, empathy, and responsibility.

In this research, the questions created refer to indicators of complex thinking abilities in aspects of decision-making using Presseisen's taxonomy. Presseinsen's taxonomy is a taxonomy that discusses thinking skills consisting of four categories in the thinking process, namely, essential thinking processes, higher-level (complex) thinking processes, epistemic thinking processes, and metacognitive processes (Presseinsen, 1985). (Presseisen, 1985) states that there are six relative indicators in decision-making: determining goals, identifying obstacles to achieving goals, identifying alternatives, analyzing alternatives, ranking alternatives, and creating the best alternative. Presseisen's taxonomy moves from more straightforward to more complex (Dietrich, 1988). There is previous research that has discussed various taxonomies of thinking, such as Marzano's taxonomy (Dewi et al., 2023), Bloom's taxonomy (Fikri et al., 2022). Stahl Murphy's taxonomy (Camila et al., 2023), and Anderson Kratwhol's taxonomy (Syahri & Ahyana, 2021). This research continues using Presseisen's taxonomy, which focuses on basic and complex thinking abilities.

Biology learning is a part of natural science. Biology learning is related to discovering and understanding nature systematically, which means not just mastering and gathering knowledge about facts, concepts, and principles (Harefa et al., 2022). Environmental change topics contain information about how the environment or nature can change over time. Environmental Change is a topic found in Phase E of class X SMA. In this topic, students are asked to participate actively in developing high-level thinking skills, especially in creating appropriate solutions to environmental problems. According to Suranata et al., (2020), the topic of environmental Change can be used for activities that support the development of high-level thinking skills in students. This aligns with the learning outcomes in environmental change material in the independent curriculum, namely that students can create solutions related to environmental Change, including local, national, or global problems.

Based on the problems above, the researcher is interested in research to develop a test instrument for students' complex thinking ability aspects of decision-making with the title "Development of a Complex Thinking Test Instrument for Decision Making on Class X Environmental Change Topics." This research aims to develop a complex thinking test instrument for decision-making aspects of Environmental Change. Hopefully, this research can be a reference for teachers in compiling test instruments to measure and train students' complex thinking abilities and see students' decision-making abilities. This research can contribute to improving education worldwide and is relevant to the challenges of the 21st century. It can also help students deal with global issues such as environmental change and promote active learning that encourages analysis, creativity and evaluation, especially in developing countries.

▪ METHOD

The research method used is research and development (R&D). The development model in this research refers to the development of the 4-D model (Thiagarajan et al., 1974). However, researchers only used three stages (3-D): the definition, design, and development. This is because research activities require in-depth analysis in order to produce a quality decision-making complex thinking test instrument, so that it can be widely used. The development of test instruments uses the stages of test instrument development proposed by Mardapi (2008) with nine stages, which researchers modified into six stages as follows: (1) compiling test specifications, (2) writing test questions, (3) reviewing test questions, (4) conducting tests try the test (5), analyze the question items, and (6) improve the test.

The population of this study was all class X students of senior high schools in Pontianak City, which had a population of 3,946 students. Then, the number of representative student samples was calculated using the Slovin formula. According to Riyanto & Hatmawan (2020) , the Slovin formula can be formulated as follows:

$$n = \frac{N}{N(e)^2 + 1}$$

Information:
n: Number of samples
N: Total population
E: Error rate in sampling (5%)

After calculating, a sample of 363 students was obtained. This student will be given the test. Sampling was conducted using a cluster sampling technique, and schools were taken randomly through a lottery. The cluster sampling technique is the expansion of the area, which is divided into five sub-districts in Pontianak so that five schools are obtained from each sub-district: SMAN 3 South Pontianak, SMAN 9 East Pontianak, SMAN 5 North Pontianak, SMAN 11 West Pontianak, and SMAN 8 Pontianak. The sample used can be seen in table 1.

Tabel 1. Cluster sampling technique

Cluster	School	Many Samples
West Pontianak District	State High School 11 Pontianak	73
East Pontianak District	State High School 9 Pontianak	72

Pontianak City District	State High School 8 Pontianak	73
North Pontianak District	State High School 5 Pontianak	72
South Pontianak District	State High School 3 Pontianak	73
Southeast Pontianak District	-	-
Total of All Samples		363

Data were collected through interviews, test instruments, and validation sheets. The data collection instruments are interview guide sheets, research instrument validation sheets, and complex thinking test sheets for aspects of decision-making. Interview guide sheet consisting of 12 questions. Research instrument validation sheet consisting of 3 assessment aspects: material aspect, construction aspect, and language aspect. The complex thinking test sheet for aspects of decision-making is prepared based on decision-making indicators which consist of 6 aspects: determine goals, identify obstacles to achieving goals, identify alternatives, analyze alternatives, rank alternatives, and choose the best alternative. Each indicator consists of two description questions, for a total of 12 description questions.

Data analysis consisted of content validity, inter-rater reliability, and Rasch analysis. The validation process was carried out by seven validators consisting of 2 lecturers Biology Education, the Faculty of Teacher Training and Education, Tanjungpura University, and five high school biology teachers. The validation sheet uses a Likert scale, which consists of 4 categories of Likert scale according to Mardapi (2008), which can be seen in Table 2.

Table 2. Likert scale

Number	Category	Score
1.	Strongly Agree	4
2.	Agree	3
3.	Disagree	2
4.	Strongly Disagree	1

The data obtained were analyzed using Aiken's V formula (1985). Content validity is considered valid if the value obtained is above the minimum value determined based on Aiken's V table. Based on the number of assessors to measure content validity, divided into four categories, the minimum standard Aiken index for this research is 0.76.

Measurement of interrater reliability was analyzed through the intraclass correlation coefficient (ICC) using the SPSS version 29 application. The inter-rater reliability analysis results are in the ICC output, which shows the Average measure value. The following are the criteria for reliability levels based on categories referring to (Perinetti, 2018), which can be seen in Table 3.

Table 3. ICC statistical criteria

ICC Value	Interpretation
<0.50	Bad
0.51 – 0.75	Enough
0.76 – 0.90	Good
0.90 – 1.00	Excellent

The validity results obtained in the good category indicate that the instrument has been able to measure the extent of the truth to be measured. Meanwhile, the reliability results obtained in the good category indicate that the instrument has been able to measure the extent to which the results are consistent with the results being measured. After the test questions are completed, the items are evaluated to analyze the questions' suitability level and difficulty level of the questions to detect bias and reliability (Cronbach's alpha, person reliability, item reliability). This analysis process uses a Rasch modeling application called Winstep. One of the advantages of the RASCH model is that it can test the suitability of people and items simultaneously; can provide a linear scale with equal intervals; can detect imprecision in the model; can produce replicable measurements; and can provide more precise estimates (Hindrasti, Sabekti, & Sarkity, 2021).

▪ RESULT AND DISSCUSSION

The define stage consists of several analysis stages: front-end analysis, topic analysis, and learning objective analysis. Interviews with biology teachers at schools conducted a front-end analysis. The results of the interviews show that the teaching and learning process still faces several problems. There are three main problems: firstly, the learning process still depends on the teacher; secondly, students need test instruments that help them to think complexly, especially in decision-making; and thirdly, students need help when given questions that address higher-level thinking abilities. The test instruments used only measure lower-level thinking abilities from understanding to application while analyzing and creating, and they still need to be used in schools. In addition, there has been no development of test instruments that address complex thinking aspects of decision-making.

Based on the interview results, the teacher suggested using the topic of environmental change. The reason is that this topic is directly related to the environmental conditions of students. Environmental change contains sub-topics: environmental change, pollution, global warming, environmental conservation, and waste recycling. Learning objectives are analyzed based on learning outcomes in phase E in the Merdeka Curriculum. The learning objectives used by teachers on the topic of environmental change are that students can identify facts about environmental changes occurring around them by presenting accurate data results, students can identify human activities that cause environmental changes, students can analyze causes and negative impacts. as well as environmental pollution efforts, students can analyze environmental conservation efforts with appropriate alternatives, and students can create solutions to overcome environmental problems.

The design stage was carried out to design the test instrument being developed. This stage consists of compiling test specifications, writing test questions, and reviewing test questions. Test specifications contain a description that shows the overall characteristics that the test instrument must have. Preparing test specifications includes determining the test objectives, compiling a grid, determining the test's form, and determining the test's length. In terms of objectives, the test used was summative in this study. Summative tests are carried out after learning ends or after environmental change material is presented. In the process of preparing the test grid, it is arranged based on the learning objectives of the environmental change material. It refers to the Presseisen taxonomy of complex thinking abilities in aspects of decision-making. The test grid is presented in matrix form,

which contains components: learning objectives, aspects of decision-making ability, material/sub-material, question indicators, and question numbers. The test developed is a written test in the form of a non-objective description with 12 questions. According to Mardapi (2008) non-objective tests can measure levels of thinking from low to high, from memorization to evaluation. Students are given 90 minutes to work on the test questions.

The test questions are written according to indicators of complex thinking aspects of decision-making. They are based on a grid of questions and topics related to environmental change that have been studied. Indicators in decision-making are determining goals, identifying alternatives, identifying obstacles to achieving goals, analyzing alternatives, ranking alternatives, and creating the best alternative. Each decision-making indicator consists of two questions, so there are 12 questions.

Examining test questions consists of content validity and inter-rater reliability. The test questions were validated by seven validators: 2 biology education lecturers at FKIP Tanjungpura University and five biology teachers. The score criteria use a Likert scale (1-4). Furthermore, they are assessed for each item on the validation sheet. Content validation was calculated using Aiken's V formula. The questions that have been prepared are reviewed for content validity and inter-rater reliability. Content validation was calculated using Aiken's V formula. The results of the content validity analysis can be seen in Table 4.

Table 4. Content validity analysis results

Decision Making Indicator	Question Number	V Aiken	Information
Determining goals	1	0.89	Valid
	7	0.95	Valid
Identifying obstacles to achieving goals	2	0.91	Valid
	3	0.94	Valid
Identifying alternatives	4	0.92	Valid
	5	0.94	Valid
Analysing alternatives	6	0.94	Valid
	10	0.93	Valid
Ranking alternatives	8	0.96	Valid
	9	0.92	Valid
Creating the best alternatives	11	0.97	Valid
	12	0.96	Valid
Average			0.94 (Valid)
Table V Aiken ($\alpha = 0.05$)			0.76 (Valid)

Based on Table 3, it is known that the total average validation for these three aspects is 0.94. These results show the value of Vcount > Vtable, namely 0.94 > 0.76, which shows that the test instrument developed is valid.

Interrater reliability calculation by the SPPS version 29 application based on the interclass correlation coefficient (ICC). The inter-rater reliability analysis results are seen from the average measure value in the ICC table. The average measure value obtained was 0.806. Perinetti (2018) states that the results obtained are reliable. The results of the inter-rater reliability analysis can be seen in Figure 1.

Interclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidance Interval		F Test with True Value			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.293[a]	.086	.711	5.143	6	54	<.001
Average Measures	.806[c]	.484	.961	5.143	6	54	<.001

Figure 1. Interrater reliability results

Validity and reliability tests are significant to ensure that the instruments used produce accurate and reliable results (Marthiani, 2024). After carrying out content validity and inter-rater reliability, improvements were made to the questions obtained from criticism and suggestions from the validator. After the test trials were carried out, the results of the test questions were analyzed using Rasch modeling, which aims to obtain information about the characteristics of the question items (Elviana, 2020). In this study, analysis was carried out to measure the questions' suitability, difficulty level, and reliability (Cronbach's alpha, person reliability, item reliability).

Item validity of a test is the measuring accuracy of an item (which is an inseparable part of the test as a totality) in measuring what should be measured through that item (Susanty, 2016). According to Maulana et al. (2023), a question item is declared valid if it meets one of the three conditions for item fit criteria. Boone & Staver, (2020) stated that the outfit mean-square, outfit Z standard, and point measure correlation values are the criteria used to see the level of item fit. Outfit mean square (MNSQ) value accepted: 0.5 < MNSQ < 1.5. Outfit Zstandard (ZSTD) value accepted: –2.0 < ZSTD < +2.0. Point Measure Correlation (Pt Measure Corr) value: 0.4 < Point Measure Corr < 0.85. The results of the item validity analysis can be seen in Figure 2.

File Edit Format View Help

TABLE 10.1 C:\Users\USER\OneDrive\Desktop\DATA H ZOU465WS.TXT Jul 29 3:01 2024
INPUT: 363 Person 12 Item REPORTED: 363 Person 12 Item 4 CATS WINSTEPS 3.73

Person: REAL SEP.: 2.41 REL.: .85 ... Item: REAL SEP.: 6.85 REL.: .98

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
8	1026	363	.04	.08	1.85	9.6	1.79	9.1	A .61	.65	32.0	56.0	E8
1	1099	358	-.58	.09	1.11	1.5	1.24	3.0	B .41	.64	53.4	58.2	E1
9	1024	358	-.03	.08	1.09	1.3	1.07	.9	C .71	.65	55.9	56.2	E9
12	898	352	.74	.08	1.09	1.2	1.07	1.0	D .68	.65	50.0	54.9	E12
2	1117	358	-.71	.09	.95	-.6	.97	-.3	E .57	.64	60.3	59.0	E2
11	940	360	.57	.08	.94	-.9	.92	-1.1	F .66	.65	60.6	55.1	E11
10	927	357	.61	.08	.93	-1.0	.91	-1.3	f .67	.65	58.8	55.0	E10
5	1000	360	.17	.08	.87	-1.9	.86	-2.1	e .70	.65	60.6	55.5	E5
7	1103	363	-.51	.09	.84	-2.3	.85	-2.2	d .67	.64	61.4	58.0	E7
3	1180	363	-1.10	.09	.84	-2.2	.79	-2.7	c .66	.62	68.6	61.1	E3
4	1035	358	-.12	.08	.74	-4.0	.73	-4.2	b .73	.65	64.0	56.2	E4
6	860	347	.94	.08	.66	-5.4	.65	-5.4	a .69	.64	70.6	55.4	E6
MEAN	1017.4	358.1	.00	.08	.99	-.4	.99	-.4			58.0	56.7	
S.D.	93.2	4.5	.61	.00	.29	3.6	.29	3.6			9.6	1.9	

TABLE 10.3 C:\Users\USER\OneDrive\Desktop\DATA H ZOU465WS.TXT Jul 29 3:01 2024
INPUT: 363 Person 12 Item REPORTED: 363 Person 12 Item 4 CATS WINSTEPS 3.73

Figure 2. Results of question suitability analysis

Based on Figure 2, it is known that of the 12 questions, 12 questions (E is the question) were accepted: E1, E2, E3, E4, E5, E6, E7, E8, E9, E10, E11, E12. Bias detection analysis of the questions was also carried out to see whether the questions developed functioned well. Question items contain bias if the probability value of the item is below 5% (Kurniawan & Andriyani, 2018). Detection of biased items or differential item functioning (DIF) is an item in the test with a different function. A question item is said to contain bias if the probability value (PROB) of the question item is below 0.05 (5%) (Sumintono & Widhiarso, 2015). The bias question detection analysis results can be seen in Figure 3.

TABLE 30.4 C:\Users\USER\OneDrive\Desktop\DATA H ZOU001WS.TXT Aug 10 20:21 2024
INPUT: 363 Person 12 Item REPORTED: 363 Person 12 Item 4 CATS WINSTEPS 3.73

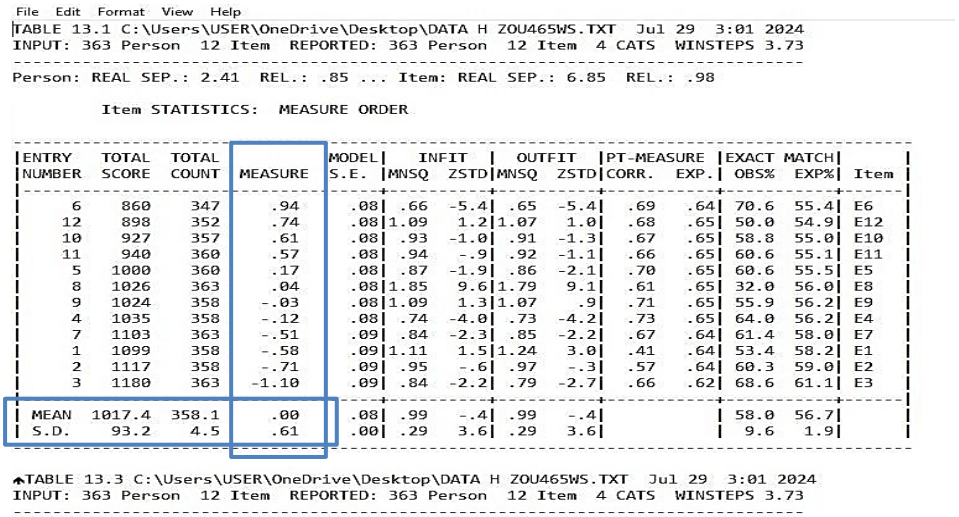
DIF class specification is: DIF=$S1W1

Person CLASSES	SUMMARY DIF CHI-SQUARE	D.F.	PROB.	BETWEEN-CLASS MEAN-SQUARE	t=ZSTD	Item Number	Name
2	.0000	1	1.0000	.0038	-1.3183	1	E1
2	.0000	1	1.0000	.0011	-1.4328	2	E2
2	.0403	1	.8408	.0299	-.9912	3	E3
2	.0000	1	1.0000	.0100	-1.1925	4	E4
2	.4439	1	.5053	.2206	-.3681	5	E5
2	.2187	1	.6400	.1085	-.6383	6	E6
2	.0820	1	.7747	.0414	-.9161	7	E7
2	.2581	1	.6114	.1282	-.5803	8	E8
2	2.2336	1	.1350	1.1163	.5507	9	E9
2	3.0874	1	.0789	1.5419	.8008	10	E10
2	.5099	1	.4752	.2542	-.3061	11	E11
2	.0000	1	1.0000	.0034	-1.3309	12	E12

Figure 3. Results of bias question detection

Figure 3 shows that 12 questions have probability values above the average probability value, namely 5% (0.05). A question item's difficulty level is obtained from the student's ability to answer the question. The difficulty level of the questions provides information regarding the difficulty level in the categories straightforward, easy, difficult, and challenging. Grouping of test items was carried out using the standard deviation (SD) value and the average logit value. The standard deviation (SD) value is 0.61, and the average logit value is 0.00. Based on these values, the challenging category is > 0.61, the difficult category is > 0.00 to < 0.61, the easy category is > -0.61 to < 0.00, and the straightforward category is < -0.61. The results of the analysis of the questions' difficulty level can be seen in Figure 4.

Based on Figure 4,, the analysis of the level of difficulty of the questions carried out on the 12 descriptive questions shows that three questions (E6, E10, E12) are in the challenging category with a percentage of 25%, three questions (E5, E8, E11) are in the category difficult with a percentage 25%, four questions (E1, E4, E7, E9) are in the easy category with a percentage 33.3%, and two questions (E2, E3) are in the straightforward category with a percentage 16, 7%. Based on these results, the question items still need consistency when viewed from the decision-making indicators. However, compare the question items with the categories (challenging, difficult, easy, and straightforward). It is



File Edit Format View Help

TABLE 13.1 C:\Users\USER\OneDrive\Desktop\DATA H ZOU465WS.TXT Jul 29 3:01 2024
INPUT: 363 Person 12 Item REPORTED: 363 Person 12 Item 4 CATS WINSTEPS 3.73

Person: REAL SEP.: 2.41 REL.: .85 ... Item: REAL SEP.: 6.85 REL.: .98

Item STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
6	860	347	.94	.08	.66	-5.4	.65	-5.4	.69	.64	70.6	55.4	E6
12	898	352	.74	.08	1.09	1.2	1.07	1.0	.68	.65	50.0	54.9	E12
10	927	357	.61	.08	.93	-1.0	.91	-1.3	.67	.65	58.8	55.0	E10
11	940	360	.57	.08	.94	-.9	.92	-1.1	.66	.65	60.6	55.1	E11
5	1000	360	.17	.08	.87	-1.9	.86	-2.1	.70	.65	60.6	55.5	E5
8	1026	363	.04	.08	1.85	9.6	1.79	9.1	.61	.65	32.0	56.0	E8
9	1024	358	-.03	.08	1.09	1.3	1.07	.9	.71	.65	55.9	56.2	E9
4	1035	358	-.12	.08	.74	-4.0	.73	-4.2	.73	.65	64.0	56.2	E4
7	1103	363	-.51	.09	.84	-2.3	.85	-2.2	.67	.64	61.4	58.0	E7
1	1099	358	-.58	.09	1.11	1.5	1.24	3.0	.41	.64	53.4	58.2	E1
2	1117	358	-.71	.09	.95	-.6	.97	-.3	.57	.64	60.3	59.0	E2
3	1180	363	-1.10	.09	.84	-2.2	.79	-2.7	.66	.62	68.6	61.1	E3
MEAN	1017.4	358.1	.00	.08	.99	-.4	.99	-.4			58.0	56.7	
S.D.	93.2	4.5	.61	.00	.29	3.6	.29	3.6			9.6	1.9	

TABLE 13.3 C:\Users\USER\OneDrive\Desktop\DATA H ZOU465WS.TXT Jul 29 3:01 2024
INPUT: 363 Person 12 Item REPORTED: 363 Person 12 Item 4 CATS WINSTEPS 3.73

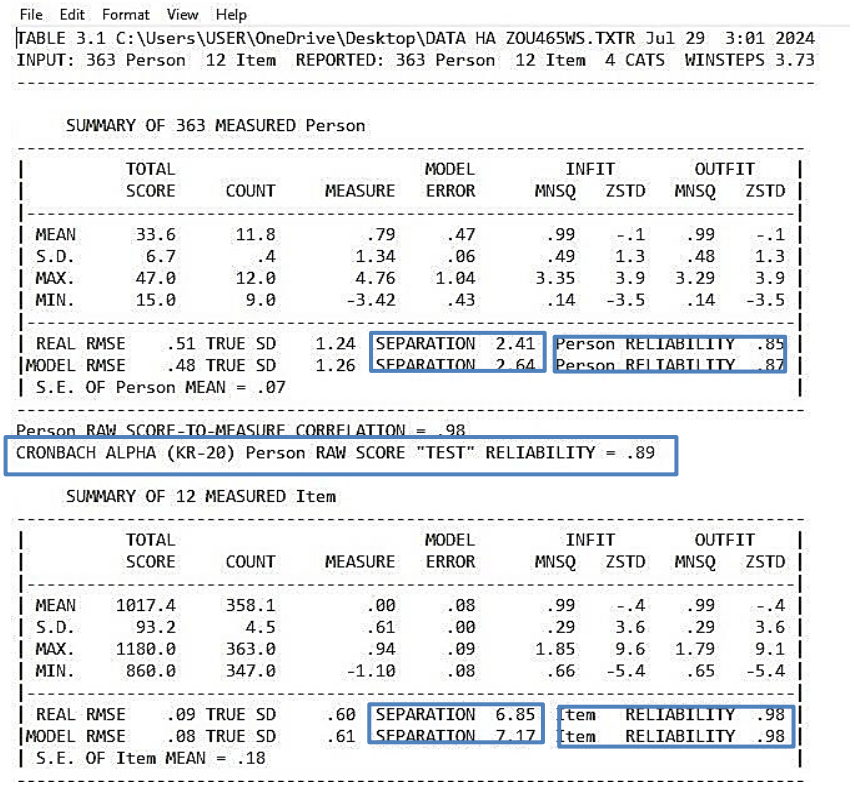
Figure 4. The result of the analysis of the questions difficulty level

balanced. The question items contain all categories of question difficulty level, which means the questions are relatively straightforward. It is in line with (Arifin, 2017), (Ardhani, 2020), (Fatimah & Alfath, 2019) that a question is said to be good if it has a proportional level of difficulty, meaning that the question is not too easy or too difficult. Test instruments that have different difficulties can collect data on differences in students who have high, medium, and low levels of decision-making ability.

Reliability can show that measurement results remain consistent if carried out twice or more on the same instrument, using the same measuring instrument (Sugiono et al., 2020). Reliability is determined to see the instrument's consistency in measurement when used repeatedly (Octaviana et al., 2022). The reliability test in this research used Rasch model analysis with the Winsteps program through the Summary Statistics table, which shows Cronbach's alpha value, person reliability, and item reliability. The results of the reliability analysis can be seen in Figure 5.

Based on Figure 5, it is known that Cronbach's alpha value is 0.89, which means that the interaction between person and item as a whole is in the excellent category. The value of person reliability is 0.85, and item reliability is 0.98. The consistency of the respondents' answers is excellent, and the quality of the items in the instrument is in the particular category. It shows that the reliability test results produce valid values.

This research has several advantages, especially in applying comprehensive analysis methods to develop test instruments focusing on complex thinking skills, especially in decision-making. The resulting test instrument assesses basic skills and more complex thinking skills, such as creation, analysis, and evaluation. This research benefits the world of education, especially in making assessment tools that can measure students' complex thinking skills. The findings of this study can help teachers better understand learners' thinking skills, teachers can identify the strengths and weaknesses of individual students in the aspect of decision-making. If learners are still weak in making decisions then teachers can design learning activities that are able to turn off this decision making. Teachers can design learning activities such as setting goals, identifying problems, analyzing, and making conclusions, so that learners are familiar with learning that trains good decision-making skills. for example, with learning strategies that can facilitate



File Edit Format View Help

TABLE 3.1 C:\Users\USER\OneDrive\Desktop\DATA HA ZOU465WS.TXTR Jul 29 3:01 2024
INPUT: 363 Person 12 Item REPORTED: 363 Person 12 Item 4 CATS WINSTEPS 3.73

SUMMARY OF 363 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	33.6	11.8	.79	.47	.99	-.1	.99	-.1
S.D.	6.7	.4	1.34	.06	.49	1.3	.48	1.3
MAX.	47.0	12.0	4.76	1.04	3.35	3.9	3.29	3.9
MIN.	15.0	9.0	-3.42	.43	.14	-3.5	.14	-3.5
REAL RMSE	.51	TRUE SD	1.24	SEPARATION	2.41	Person RELIABILITY		.85
MODEL RMSE	.48	TRUE SD	1.26	SEPARATION	2.64	Person RELIABILITY		.87
S.E. OF Person MEAN = .07								

Person RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .89

SUMMARY OF 12 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	1017.4	358.1	.00	.08	.99	-.4	.99	-.4
S.D.	93.2	4.5	.61	.00	.29	3.6	.29	3.6
MAX.	1180.0	363.0	.94	.09	1.85	9.6	1.79	9.1
MIN.	860.0	347.0	-1.10	.08	.66	-5.4	.65	-5.4
REAL RMSE	.09	TRUE SD	.60	SEPARATION	6.85	Item RELIABILITY		.98
MODEL RMSE	.08	TRUE SD	.61	SEPARATION	7.17	Item RELIABILITY		.98
S.E. OF Item MEAN = .18								

Figure 5. The results of the reliability analysis (*summary statistik*)

training this skill. In addition, this test instrument can be used by teachers to determine the extent to which students are able to identify goals, analyze various options, and make appropriate decisions in complex situations. The limitations of this study include testing the instrument only on one material, namely environmental changes. In addition, the development stage only uses 3-D due to limited cost, time, and research resources. For future research, it is recommended that the developed instrument be tested on other materials or subjects and continue the research until the dissemination stage. Based on previous research conducted by Utami, et al (2023) on the analysis of decision-making skills on renewable energy material, the results of decision-making skills are still relatively low. Therefore, it is necessary to increase high-level thinking, especially in decision-making skills.

▪ CONCLUSION

Based on the research results, the test instrument developed was in the form of 12 essay questions containing aspects of decision-making and environmental change topics. The test instrument is a summative test which is tested after the topic of environmental change is taught. The time to complete the test questions is 90 minutes or one and a half hours. Before the test questions were tested, a review was carried out; content validation analysis showed that the complex thinking test instrument for aspects of decision-making that was developed was valid because V-count =0.94 > (V-table =0.76). Interrater reliability analysis obtained a value of 0.806, included in the good category. Analysis of the difficulty level shows that three questions are challenging, three are difficult, four are easy, and two are straightforward. Reliability analysis also uses the Rasch model, which

shows a Cronbach's alpha value of 0.89, meaning that the interaction between the person and the items or questions is excellent. Furthermore, the person reliability value is 0.85, meaning the consistency of the students' answers is exemplary, and item reliability is 0.98, meaning the quality of the question items is special.

▪ REFERENCES

Afrita, M., & Rahmawati, D. 2020. *Validitas instrumen tes berpikir tingkat tinggi (hots) pada materi sistem respirasi di kelas XI SMA.* Mangifera Edu, 4(2), 129–142. https://doi.org/10.31943/mangiferaedu.v4i2.83

Akmala, N. F., Suana, W., & Sesunan, F. 2019. *Analisis kemampuan berpikir tingkat tinggi siswa sma pada materi hukum newton tentang gerak.* Titian Ilmu: Jurnal Ilmiah Multi Sciences, 11(2), 67–72. https://doi.org/10.30599/jti.v11i2.472

Ardhani, Y. 2020. *Kualitas butir soal penilaian akhir tahun mata pelajaran teknologi dasar otomotif kelas x teknik kendaraan ringan otomotif di Smk Muhammadiyah Gamping.* Jurnal Pendidikan Vokasi Otomotif, 3(1), 85–94. https://doi.org/10.21831/jpvo.v3i1.34917

Arifin, Z. 2017. *Kriteria instrumen dalam suatu penelitian.* Jurnal Theorems (the Original Research of Mathematics), 2(1), 28–36. https://doi.org/10.31949/th.v2i1.571

Boone, W. J., & Staver, J. R. 2020. Advances in rasch analyses in the human sciences. advances in rasch analyses in the human sciences. https://doi.org/10.1007/978-3-030-43420-5

Camila, C. G., Afandi, A., Tenriawaru, A. B., Artika, W., & Siregar, N. 2023. Development of higher order thinking skill questions using Stahl and Murphy's taxonomy on excretion system topic. Assimilation: Indonesian Journal of Biology Education, 6(2), 97–108. https://doi.org/10.17509/aijbe.v6i2.60632

Dietrich, C. 1988. Critical thinking in the college currtculum: somev/here or nov/here? In What Socrates Began: An Examination of the lntellect (Libby G. C, pp. 20–24). the Univenity of Southern Maine Portland, Main. https://digitalcommons.usm.maine.edu/cgi/viewcontent.cgi?article=1237&context=facbooks#page=21

Elan, E., Sumardi, S., & Juandi, A. S. 2022. *Penyusunan instrumen penelitian tindakan kelas dalam upaya peningkatakan keterampilan sosial.* Jurnal Paud Agapedia, 6(1), 91–98. https://doi.org/10.17509/jpa.v6i1.51339

Elviana. 2020. *Analisis butir soal evaluasi pembelajaran pendidikan agama islam menggunakan program anates.* Jurnal MUDARRISUNA, 10(2), 58–74. https://jurnal.ar-raniry.ac.id/index.php/mudarrisuna/article/view/7839

Fatimah, Laela Umi, & Alfath, K. 2019. *Analisis kesukaran soal, daya pembeda dan fungsi distraktor.* Sustainability (Switzerland), 11(1), 1–14. http://scioteca.caf.com/bitstream/handle/123456789/1091/RED2017-Eng-8ene.pdf?sequence=12&isAllowed=y%0Ahttp://dx.doi.org/10.1016/j.regsciurbeco.2008.06.005%0Ahttps://www.researchgate.net/publication/305320484_Sistem_Pembetungan_Terpusat_Strategi_Melestari

Fikri, A. M. K., Sudarti, S., & Handayani, R. D. 2022. *Analisis deskriptif kemampuan berpikir tingkat tinggi (hots) siswa ma unggulan nurul iman pokok bahasan suhu dan kalor dengan menggunakan taksonomi bloom.* Jurnal Pendidikan Mipa, 12(2), 214–219. https://doi.org/10.37630/jpm.v12i2.581

Harefa, M., Lase, N. K., & Zega, N. A. 2022. *Deskripsi minat dan motivasi belajar siswa pada pembelajaran biologi.* 1(2), 381–389. https://doi.org/10.56248/educativo.v1i2.65

Hindrasti, N. E. K., Sabekti, A. W., & Sarkity, D. (2021). *Pelatihan menyusun soal kemampuan berpikir kritis dan analisis menggunakan model rasch bagi guru ipa.* Reswara: Jurnal Pengabdian Kepada Masyarakat, 2(2), 212-219. https://doi.org/10.46576/rjpkm.v2i2.1066

Kurniawan, & Andriyani, K. D. K. 2018. *Analisis soal pilihan ganda dengan rasch model.* Jurnal Statistika, 6(1), 34–39. https://doi.org/10.26714/jsunimus.6.1.2018.%25p

Magdalena, I., Syariah, E. N., Mahromiyati, M., & Nurkamilah, S. 2021. *Analisis instrumen tes sebagai alat evaluasi pada mata pelajaran sbdp siswa kelas ii sdn duri kosambi 06 pagi.* Jurnal Pendidikan Dan Ilmu Sosial, 3(2), 276–287. https://ejournal.stitpn.ac.id/index.php/nusantara

Mardapi, D. 2008. *Teknik penyusunan instrumen tes dan non tes.* mitra cedikia press.

Marthiani, I. 2024. *Uji validitas dan reliabilitas instrumen penelitian kuantitatif.* Jurnal Ilmiah Kependidikan, 2(2), 17–23. https://doi.org/10.61132/yudistira.v2i2.727

Maulana, AK, & Rochintaniawati, D. (2021). Analysis of decision-making skills of grade XI Students of SMAN 1 Cihaurbeuti. ISEJ: Indonesian Science Education Journal, 2(2), 83–89. DOI:10.4135/9781506307633.n706 Jurnal Ilmiah Kependidikan, 2(2), 17–23. https://doi.org/https://doi.org/10.61132/yudistira.v2i2.727

Mufidah, S., & Wijaya, A. 2017. *Pengembangan perangkat pembelajaran matematika realistik pada materi aritmatika sosial untuk meningkatkan kemampuan berpikir tingkat tinggi siswa smp kelas VII.* 6, 11–18. https://journal.student.uny.ac.id/index.php/jpm/article/download/6970/6695

Nasution, H. fadilah. 2016. *Instrumen penelitian dan urgensinya dalam penelitian kuantitatif.* Sustainability (Switzerland), 4(1), 59–75. https://doi.org/10.24952/masharif.v4i1.721

Octaviana, R. I., Anggara, M. B., Jamilah, R., Darmana, A., & Suyanti, R. D. 2022. *Analisis item soal kimia SMA menggunakan rasch model.* Orbital: Jurnal Pendidikan Kimia, 6(1), 26–37. https://doi.org/10.19109/ojpk.v6i1.12248

Perinetti, G. 2018. StaTips Part IV: Selection, interpretation and reporting of the intraclass correlation coefficient. South European Journal of Orthodontics and Dentofacial Research, 5(1). https://doi.org/10.5937/sejodr5-17434

Presseisen, B. . 1985. Thinking skills throughout the curriculum: a conceptual design. Research for Better Schools, Inc.

Rahmasari, Y., Noviekayati, I., & Pratitis, N. T. 2023. Self-Determination and conformity with student career decision making , how are they related? International Journal of Social and Management Studies (Ijosmas), 4(6), 27–32. https://doi.org/10.5555/ijosmas.v4i6.376

Riyanto, S., & Hatmawan, Aglis, A. 2020. *Metode riset penelitian kuantitatif penelitian di bidang manajemen, teknik, pendidikan dan eksperimen.* CV Budi Utama.

Sanjaya, M. S. M., Wahidin, & Maryuningsih, Y. 2019. *Penerapan pembelajaran biologi berbasis entrepreneurship pada materi.* Jurnal Ilmu Alam Indonesia, 2(1), 21–35. https://syekhnurjati.ac.id/jurnal/index.php/jia/article/view/6272

Sugiono, Noerdjanah, & Wahyu, A. 2020. *Uji validitas dan reliabilitas alat ukur sg posture evaluation.* Jurnal Keterapian Fisika, 5(1), 55–61. https://doi.org/10.37341/jkf.v5i1.167

Sumintono, B., & Widhiarso, W. 2015. *Aplikasi pemodelan rasch pada assessment pendidikan (issue september).* Trim Komunikata Publishing House. http://eprints.um.edu.my/id/eprint/14228

Suranata, K., Apriliana, I. P. A., & Ifdil, I. 2020. The effect of problem-solving training to improve student's critical thinking and decision-making skills: racked analysis. Acta Counseling and Humanities, 1(1), 1–9. https://doi.org/10.46637/ach.v1i1.6

Suratno, S., Kamid, K., & Sinabang, Y. 2020. *Pengaruh penerapan model pembelajaran problem based learning (PBL) terhadap kemampuan berpikir tingkat tinggi (HOTS) di tinjau dari motivasi belajar.* Jurnal Manajemen Pendidikan Dan IImu Sosial, 1(2), 506–515. https://doi.org/10.38035/JMPIS

Syahri, andi alim, & Ahyana, N. 2021. *Analisis kemampuan berpikir tingkat tinggi menurut teori anderson dan krathwohl.* 1(1), 41–52. https://doi.org/10.51574/jrip.v1i1.16

Thiagarajan, Sivasailam, & Others. 1974. Instructional development for training teachers of exceptional children: a sourcebook. Indiana Univ., Bloomington. Center for Innovation in (Issue Mc). Indiana University. https://files.eric.ed.gov/fulltext/ED090725.pdf

Utami, A. C., Khoiri, N., Saefan, J., & Ristanto, S. (2023). *Analisis keterampilan pengambilan keputusan pada pemecahan masalah fisika peserta didik kelas X SMA N 1 Mranggen.* Jurnal Inovasi Pembelajaran di Sekolah, 4(2), 721-727. https://doi.org/10.51874/jips.v4i2.176