Dissemination of test instruments as product of the development research to measure the problem-solving ability of class X students by online in the pandemic period

Dissemination of test instruments as product of the development research to measure the problem-solving ability of class X students by online in the pandemic period

E Susilaningsih*, M Nuswowati, N Aprilia, and A Luthfiyah

Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

*Corresponding author: endang.arkan@gmail.com

Abstract. Assessment of online problem-solving abilities that are integrated on knowledge test questions to measure the cognitive aspects of students is very necessary during a pandemic. This study aims to disseminate test instruments, to measure students' problem-solving abilities. This research uses descriptive quantitative method, with the design of test instrument preparation, the implementation of test, and the data analysis of test result. The research subjects were students of class X, totaling 75 students. Data collection was carried out by tests and questionnaire sheets. Data analysis used the IRT Rasch model to estimate separation item reliability, item fit, wright map, item measure, person fit, person measure and DIF items. The results showed the profile of students' problem-solving abilities with sufficient criteria was 97% namely 73 of 75 students, with 16% classical completeness, namely 12 of 75 students. The results of the item analysis using the Rasch model showed the overall reliability of the test was 0.73, a good criterion. The results of the student response questionnaire recapitulation were as many as 95% of students gave positive responses to the developed test instruments. This test instrument is able to measure students' problem-solving abilities and can help implement learning outcomes tests during a pandemic.

1. Introduction

Students need problem-solving skills. In preparation for the challenges of the 21st century, students are required to have problem-solving skills. Problem-solving skills are needed so that students can compete globally in this century [1]. One of the solutions used in realizing 21st century skills balanced with the implementation of the 2013 curriculum is to develop an assessment of problem-solving abilities. Assessment of problem-solving abilities was developed to facilitate the teacher in the learning process to measure all student competencies [2].

Problem-solving is one of the competencies that students must possess in learning chemistry in high school [3]. Problem-solving is an individual attempt to use the knowledge and skills they have to find solutions to a problem or find new situations that have not been previously known. Problem-solving skills require complex thinking skills or higher order thinking [4]. In times of pandemics like this, online problem-solving skills assessments are needed. Therefore, researchers will develop an integrated online problem-solving ability assessment on knowledge test questions to measure students' cognitive aspects.

The test is an instrument that can be used to measure problem-solving abilities. Problem-solving abilities include the ability to analyze, interpret, reason, predict, evaluate and ponder. This means, the problem-solving ability assessment test questions must include questions that are challenging for

students to be able to do analysis, interpretation, reasoning, prediction and evaluation in problem-solving. The test has two functions. First, as a measuring and assessment tool learning outcomes such as problem-solving abilities. Second, it can be used to practice problem-solving skills [5].

The problem-solving skills assessment instrument is considered valid if the instrument can measure students' problem-solving skills. Meanwhile, instrument validation is also needed, because instrument validation can show the appropriateness of the assessment instrument function [6]. In this study, the developed test instrument will be analyzed using Item Response Theory (IRT) with Rasch Model. Analysis using the Rasch model was used to help further analyze the quality of each item to identify the ability of each student based on gender [7]. The analysis carried out can help teachers find out the difficulties of students with chemical concepts.

2. Methods

This research is a descriptive quantitative study, to distribute test instruments, product development research (dissemination stage), measure the problem-solving abilities of class X students, and get a description of the quality of the test through the characteristics of the Rasch model test. The design includes preparing test instruments, implementing test questions, and analyzing test result data.

The Rasch model emphasizes that every student has the same opportunity to answer the questions correctly and at the same time the questions have different levels of difficulty [8]. This is what Rasch calls person logit and item logit. The data was collected using tests and student response questionnaire sheets. The research subjects were students of class X, totaling 75 students. Quantitative data analysis was carried out through the Rasch model of the IRT approach with the ministep program's help.

The Rasch model can calculate the score of each respondent in the form of interval data. with this interval data, correct information can be obtained, for example, on which questions many students fail so that improvements can be made. Rasch developed a measurement model that determines the relationship between student ability levels and item difficulty levels. Measuring tools that provide information about a person's position in the attributes measured by a good measuring instrument will ensure valid and reliable results to accurately measure students' abilities [8].

Data analysis used the Item Response Theory Rasch model to estimate separation item reliability, item fit order, wright map, item measure, person fit order, person measure and detection of the biased item (DIF items) and questionnaire analysis of student responses to the test instrument. IRT is used as a substitute for Classical Test Theory (CTT), which has the weakness of dependence on tests, which means that an individual's ability is influenced by the characteristics of the items in a test, and the ability of the test taker influences vice versa, the characteristics of the items in CTT. In contrast to CTT, which focuses on the scores obtained, IRT does not depend on a specific sample of items or people selected in the test (item and person), so the measurement is more precise [9]. While the reliability analysis using the Rasch model produces item reliability, person reliability, and Cronbach Alpha (item-person reliability). This is because the Rasch model's reliability analysis is more accurate than the Cronbach Alpha test alone [10].

The Rasch model analysis can detect individuals whose response patterns do not match and the items are invalid (outliers or misfits). According to Boone et al. (2014) [11], there are three criteria used to check the suitability of items that are not suitable (outliers or misfits) and individuals whose response patterns are not suitable (not fit), namely the accepted Outfit Mean Square (MNSQ) value: 0.5 <MNSQ <1.5; Accepted Z-Standard Outfit (ZSTD) values: -2.0 <ZSTD <+2.0; and the accepted value of Point Measure Correlation (Pt Measure Corr): 0.4 <Pt Measure Corr <0.85. If the items on the three criteria are not fulfilled, it can be ascertained that the problem item is not good enough that it needs to be repaired or replaced [12]. This will ensure that the level of understanding of students is indeed tested through appropriate and quality items.

3. Result and Discussion

3.1. Instrument reliability

The value of reliability in Rasch modeling is indicated by the value of person reliability, item reliability and Cronbach's alpha value as overall reliability. Reliability data were analyzed using the Rasch model through Summary Statistics.

The instrument's summary statistical results show that-the person reliability value is 0.72 and 0.75 with sufficient category and the item reliability value is 0.95 with a very good category. It can be concluded that the consistency of the answers from the students was sufficient, but the quality of the items in the reliability aspect of the instrument was excellent. As the Cronbach alpha value (overall reliability) obtained was 0.73. This value falls into the "good" criteria. This means that the instrument developed has a good overall reliability coefficient.

3.2. Item fit order

The item fit level determines the validity of each item. Fit items were analyzed using the Rasch model through Item Fit Order. The results of the fit item analysis are presented in Figure 1.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	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
8	37	75	.19	.26	1.17	1.86	1.68	3.65	A .25	.42	67.6	67.0	S8
13	22	75	1.28	.29	1.16	1.01	1.33	1.52	B .36	.50	78.4	79.0	S13
10	39	75	.06	.26	1.29	3.28	1.25	1.49	C .20	.41	41.9	65.8	S10
5	56	75	-1.11	.28	1.04	.37	1.22	.76	D .25	.30	75.7	74.5	S5
11	43	75	-.20	.26	1.12	1.45	1.06	.38	E .31	.39	58.1	65.6	S11
9	49	75	-.60	.26	1.05	.64	1.10	.50	F .30	.35	66.2	68.1	S9
1	69	75	-2.58	.43	1.05	.27	.94	.12	G .13	.17	91.9	91.9	S1
3	25	75	1.04	.28	1.03	.24	.93	-.36	H .48	.48	71.6	76.3	S3
7	55	75	-1.03	.28	1.01	.10	.92	-.19	I .31	.31	71.6	73.5	S7
14	47	75	-.46	.26	1.01	.18	.92	-.32	J .36	.36	63.5	66.8	S14
2	70	75	-2.78	.47	.98	.08	.89	.08	j .18	.15	93.2	93.3	S2
15	32	75	.53	.26	.90	-.93	.98	-.09	i .51	.45	75.7	70.3	S15
12	22	75	1.28	.29	.96	-.22	.93	-.26	h .53	.50	81.1	79.0	S12
19	28	75	.81	.27	.95	-.36	.92	-.45	g .51	.47	75.7	73.7	S19
20	50	75	-.67	.26	.91	-.97	.83	-.67	f .41	.34	75.7	68.9	S20
18	61	75	-1.54	.31	.90	-.55	.74	-.59	e .34	.26	81.1	81.1	S18
4	24	75	1.12	.28	.86	-.89	.88	-.63	d .58	.49	81.1	77.2	S4
17	17	75	1.76	.32	.87	-.64	.64	-1.43	c .63	.51	83.8	83.2	S17
16	28	75	.81	.27	.85	-1.21	.80	-1.23	b .58	.47	78.4	73.7	S16
6	14	75	2.09	.35	.74	-1.18	.57	-1.44	a .69	.51	87.8	85.6	S6
MEAN	39.4	75.0	.00	.30	.99	.1	.98	.0			75.0	75.7	
P.SD	16.7	.0	1.32	.06	.13	1.1	.24	1.1			11.5	8.0	

Figure 1. Item fit order

The fit item explains whether the item is functioning normally to take measurements or not. If an item is not fit, it can be indicated that there is a misconception among students. This information helps teachers improve their teaching quality so that misconceptions can be avoided when teaching it again.

Based on Figure 1, the fit order item analysis results show that the items with item S8, tend not to fit (outliers). This is because item S8 does not meet the three criteria according to Boone et al. (2014) [11], namely MNSQ, ZSTD and PT MEASURE CORR, because the logit value is outside the criteria. This indicates that the item does not function normally in making measurements. Therefore item number 8 must be revised or replaced. Meanwhile, some of the other items only did not meet one criterion, so that several other items fell into the category of fit items and did not need to be revised.

3.3. Person-Item Map (Wright Map)

The Wright Map describes the distribution of student or respondent abilities and the distribution of the difficulty levels of the questions on the same scale. This Wright Map Analysis provides invaluable

information for teachers in identifying student ability. Wright maps are analyzed using the Rasch model through Variable Maps will produce a wright map. The wright map can be seen in Figure 2.

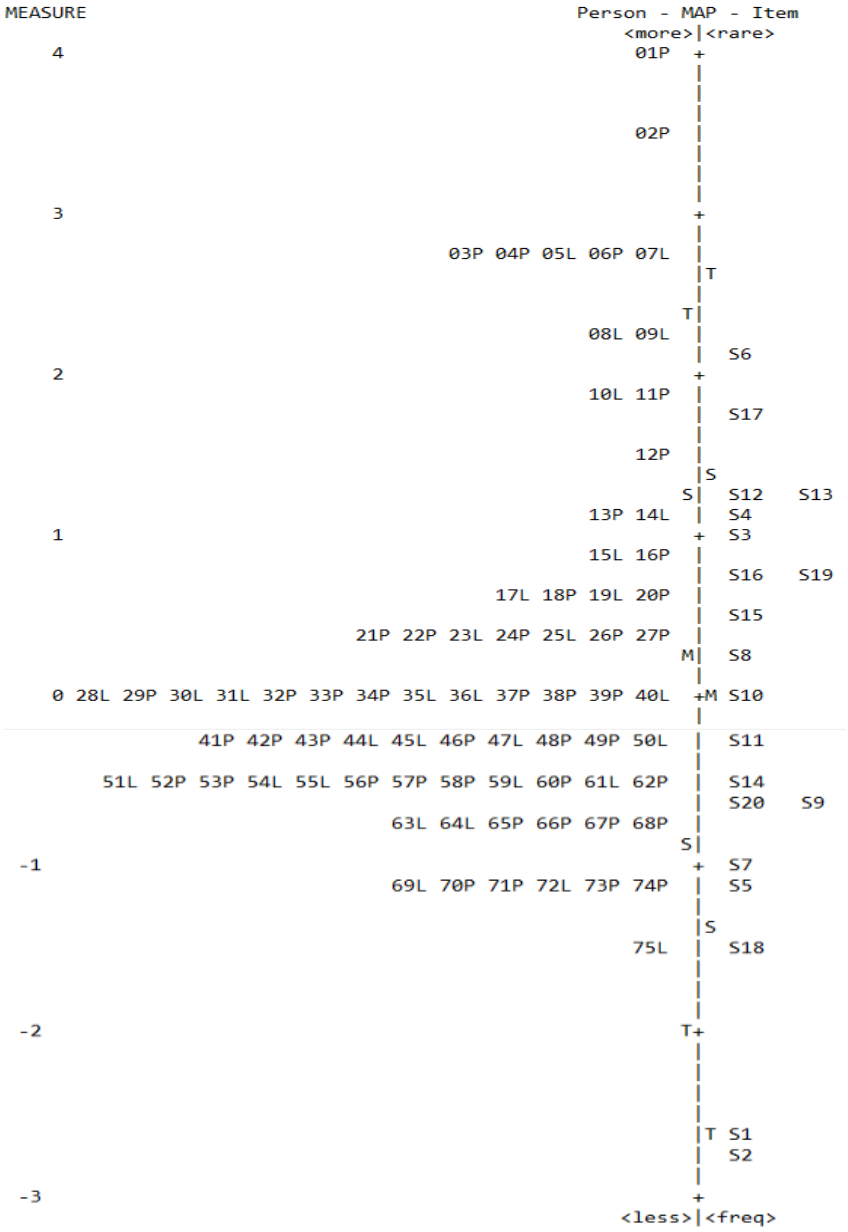

Figure 2. Wright map

The left part is the distribution of student ability or student ability while the right side of the wright map above is the distribution of the difficulty level of the items. Based on Figure 2, it can be seen that the students with the highest ability are 01P, 02P, 03P, 04P, 05L, 06P, and 07L. The student is outside the boundary of two standard deviations (T) indicating different high intelligence (outliers). If based on the logit value obtained, the seven students' logit value is more than +2.75 logit. Of course, this value is greater than the logit value of the problem with the highest difficulty level, namely the Q6 item which has a logit value of less than +2.09 logit. This indicates that, almost all the questions can be done correctly by the seven students, and the probability of all students doing the Q6 correctly is very small.

Meanwhile, the students with the lowest ability were 75L students with a logit value of less than -1 logit, but still fell within the two standard deviation (T) limit.

Based on the map above, it can also be seen that the item with the highest difficulty level is the item Q6, while the item with the lowest difficulty level or the easiest item is Q2. Problem Q2 is outside the limit of two standard deviations (T). The probability of all students to do Q6 questions correctly is very small, on the contrary, almost all students can do Q2 questions correctly.

3.4. Item Measure

A high logit value indicates the highest level of problem difficulty. The level of difficulty of the items in this study were grouped using the logit value of each item. This study categorizes the difficulty level of the questions into 4 categories based on the logit value. The difficulty level of the items is categorized by combining the standard deviation value with the logit value in the measure column. The classification of the difficulty level of the items based on the Rasch model analysis can be seen in Table 1. The Standard Deviation (SD) value in this test is 1.32.

Table 1. Classification of the difficulty level of the items based on the Rasch model analysis

Item Difficulty Level	Measure Value	Item Questions
Very Difficult	Greater than 1.32	6, 17
Difficult	0.00 – 1.32	12, 13, 4, 3, 16, 19, 15, 8, 10
Moderate	-1.32 – 0.00	11, 14, 9, 20
Easy	Less than 1.32	7, 5, 18, 1, 2

Based on Table 1, it can be seen that, of the total item items, the most questions were questions with a difficult category. Based on the results of the output measure item, question Q6 the logit value is +2.09 and for S17 the logit value is +1.76 indicating that the question is included in the very difficult category to do as evidenced by 14 students who are able to answer correctly for Q6 questions and 17 students who answered correctly for question Q17. Meanwhile, items that have the same logit value mean that the difficulty level of the items is not different.

3.5. Person Measure

Person measure analysis will provide information about individual ability, namely to identify the level of student ability to answer questions. The level of individual ability in the Rasch model analysis can be classified based on the value in the measure column with the standard deviation value.

The Standard Deviation (SD) value in this test is 1.21. Furthermore, the starting point for determining the ability of this student is from the average logit person value, the average logit person value in this test is 0.28. From the logit value the students are then grouped into 4 categories of ability, namely very high, high, medium and low. The following in Table 2 describes the criteria for grouping students' ability.

Table 2. Criteria for grouping students' ability

Ability Level	Measure Value	Student Serial Number	Number of Person
Very High	Greater than 1.21	1-12	12
High	0,28 – 1.21	13-27	15
Moderate	-1.21 – 0.28	28-74	47
Low	Less than 1.21	75	1

Based on Table 2, it can be seen that, of all students, the most ability is in the medium category. It can also be concluded that students with very high ability to solve questions were student number 1 who had a logit value of +4.83 logit, while students with low ability to solve questions were student number 75 with a logit score of -1.44 logit. This information can also be related to the logit value on the measure item, namely how much the student's ability to answer the items. Student number 1 has a logit value of +4.83 logit, while the very difficult item Q6 has a logit value of +2.09 logit. Based on the logit value data, it is clear that the logit value of the most difficult item, namely the Q6 question, is actually lower

than the logit value of students who have the ability to do very high questions. Therefore, student number 1 can easily answer item Q6 correctly.

3.6. Person Fit Order

The person fit order analysis in the Rasch model analysis serves to provide information by detecting if there are individuals who have inappropriate or different responses. A different response pattern is the mismatch of the answers given based on their ability compared to the ideal model. Individual suitability is said to be fit if it meets three criteria according to Boone *et al*., (2014) [11], namely the value of outfit means-square (MNSQ), outfit z-standard (ZSTD), and point measure correlation (PT MEASURE CORR). Person fit at this stage is analyzed using the Rasch model through Person Fit Order. A snippet of the Output Table is presented in Figure 3.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Person
19	12	20	.60	.52	1.12	.59	2.21	2.25	A .31	.48	70.0	72.4	19LD
8	17	20	2.23	.67	1.32	.86	2.01	1.10	B .05	.32	85.0	85.0	08LK
20	12	20	.60	.52	1.83	3.14	2.00	1.95	C-.08	.48	40.0	72.4	20PD
72	6	20	-1.10	.57	1.60	1.85	1.89	1.62	D .11	.51	60.0	77.7	72LK
71	6	20	-1.10	.57	1.57	1.77	1.68	1.33	E .15	.51	60.0	77.7	71PD
37	10	20	.06	.52	1.32	1.36	1.66	1.66	F .27	.51	60.0	72.5	37PD
48	9	20	-.21	.53	1.37	1.51	1.64	1.65	G .26	.51	55.0	73.1	48PK
64	7	20	-.78	.55	1.52	1.78	1.61	1.37	H .19	.52	60.0	75.6	64LK
2	19	20	3.56	1.05	1.14	.45	1.51	.78	I .03	.19	95.0	95.0	02PK
40	10	20	.06	.52	1.47	1.91	1.39	1.10	J .22	.51	50.0	72.5	40LK
50	9	20	-.21	.53	1.23	1.00	1.47	1.28	K .34	.51	65.0	73.1	50LD

Figure 3. Snippet of person fit order

Based on Figure 3, it can be seen that 19LD respondents have an unfit response pattern, because they do not meet the three criteria. Meanwhile, for other respondents, for example respondents 08LK and 20PD did not meet the MNSQ and PT MEASURE CORR criteria, but the ZSTD criteria were met. In this case, the ZSTD outfit criteria become a benchmark in the person fit analysis. If a person is found with a ZSTD outfit value that only meets the criteria while the other 2 criteria are not met, then that person is considered to still have a fit response pattern.

This unusual response pattern information can be seen in more detail by looking at the scalograms contained in the Ministep software. These scalograms can also be called the Guttman Matrix. Through these scalograms, the causes of student response patterns are not fit or ideal. Such as guesswork, indicated cheating or not being thorough. The pattern can be adjusted from the logit measure values found in the person measure analysis and measure items.

19LD students have a logit ability score of +0.6 logit which is included in the category of students with high ability. In the 19LD scalogram, students could correctly answer question Q4. Where Q4 has a logit value of +1.12 logit. This logit value is higher than the student's logitability value. However, the easiest question, namely Q2, which has a logit value lower than the logit value of the student's ability, which is -2.78 logit, cannot be answered correctly by the student. This indicates that, there is a possibility that students can answer question Q4 correctly due to lucky guess. Meanwhile, in answering Q2 questions, students indicated being careless.

3.7. Detection of Bias (Item DIF)

The items are said to contain bias if the probability value (PROB) of the items is below 0.05 (5%). PROB values can be seen in the PROB column in Output Table Item DIF. The results show that all the items used are unbiased, because the probability value (PROB) is not below 0.05. Thus, these results confirm that no individual with certain characteristics benefits more from other characteristics. All items can be worked out equally, both men and women.

3.8. Troubleshooting Profiles

The profile regarding this collaborative problem-solving ability is obtained from the combined score of the problem-solving test score and the score from the observation sheet. Based on the results of the analysis of the problem-solving ability profile, it shows that the ability is in the sufficient and inadequate category, namely 73 students from a total of 75 students, 97% in the sufficient category, while the poor category with a percentage of 3%.

3.9. Classical Completeness

Student success in a lesson is said to have completed learning (classical completeness) when indicated by a minimum completeness percentage of 75% [13]. The results obtained show that students who complete only 19%, that is, 14 out of 75 students. Based on the results of classical completeness which is only 19%, of course this greatly influences the results of the research, especially on students' problem-solving skills. The more students who complete (at least 75%), the students' problem-solving skills will increase to the good category. This small classical completeness result is related to the logit value of the student's ability and the items' logit value. The items developed have a high level of difficulty and good reliability, so they are classified as difficult questions, but the answer students' ability or consistency is low. This means that only highly capable students are able to answer all the questions correctly so that only a few students fall into the complete category. This is what causes students' low classical completeness.

3.10. Student Response Questionnaire

As many as 36%, namely, 27 of 75 students gave very agreeable responses to the developed test instrument, 59% of the other students, namely 44 of 75 students, gave agreed responses, and only 5% of students, namely 4 out of 75 students who gave responses disagree with the test instrument that was developed. Based on the results of the student response questionnaire recapitulation, of course, it can be concluded that as many as 95% of students gave positive responses to the developed test instruments.

4. Conclusion

Based on the study results, it can be concluded that the developed test instrument is valid and reliable using the Rasch model analysis with the reliability value obtained of 0.73, with good criteria. Furthermore, of the 20 items developed, only item number 8 was not fit or an outlier. Some students had an inappropriate response pattern, namely the 19LD student, which indicated that the student was guessing and was not careful in answering. All items were not detected by bias, which means that all items did not profitable male nor female students. The profile of students' problem-solving abilities with sufficient category was 97%, namely 73 out of 75 students, and 3% in the poor category. Classical completeness of students was 19%, that is, 14 of the total 75 students completed. Based on the recapitulation results of the student response questionnaire, 95% of students responded positively to the test instrument developed. This test instrument is able to measure students' problem-solving abilities and can help carry out learning outcomes tests during a pandemic.

5. Acknowledgment

Thank you to LPPM UNNES for researching the source of funds for DIPA PNBP UNNES 2020. Budget Implementation List (DIPA) Semarang State University Number: SP DIPA-023.17.2.677507/2020, December 27, 2019, in accordance with Research Development Assignment Agreement Letter UNNES DIPA Fund for 2020 114.23.4/UN37/PPK.3.1/2020, April 23, 2020.

References

[1] Permatasari A K, Istiyono E and Kuswanto H 2019 *Int. J. Educ. Res. Rev.* **4** 358
[2] Tambunan H 2019 *Int. Electron. J. Math. Educ.* **14** 293
[3] Roheni, Herman T and Jupri A 2017 *J. Phys. Conf. Ser.* **895** 012079
[4] Widiasih, Permanasari A, Riandi and Damayanti T 2018 *J. Phys. Conf. Ser.* **1013** 012081

[5] Annisah S, Zulela, Boeriswati E, Wildaniati Y and Supriatin A 2020 *Int. J. Adv. Sci. Technol.* **29** 1483
[6] Khotimah R P and Masduki 2016 *J. Res. Adv. Math. Edu.* **1** 1
[7] Planinic M, Boone W J, Susac A and Ivanjek L 2019 *Phys. Rev. Phys. Educ. Res.* **15** 020111
[8] Susongko P 2016 *J. Pendidik. IPA Indones.* **5** 268
[9] Rosa S J, Fitri A R and Agung I M 2019 *Humanitas.* **16** 54
[10] Yasin S N T M, Yunus M F M and Ismail I 2018 *J. Couns. Educ. Technol.* **1** 22
[11] Boone W J, Staver J R and Yale M S 2014 *Rasch Analysis in the Human Sciences* (Dordrecht: Springer)
[12] Sumintono B and Widhiarso W 2015 *Aplikasi Permodelan Rasch Pada Assessment Pendidikan* (Cimahi: Trim Komunikata)
[13] Savitri W R, Susilaningsih E, and Harjono 2019 *J. Inov. Pendidik. Kim.* **13** 2395