JESE 5 (1) (2025)

Journal of Environmental and Science Education

http://journal.unnes.ac.id/sju/index.php/jese

Development of a Critical Thinking Assessment Instrument for Earth Changes in Elementary School Science Lessons

Nisa Us Sa'idah [1*], Eko Handoyo [1], Supriyadi [1], Woro Sumarni [1]

Basic Education, Postgraduate, Universitas Negeri Semarang, Semarang, Indonesia 50233

DOI: 10.15294/jese.v5i1.6297

Article Info

Received 21 August 2024
Accepted 20 November 2024
Published 28 April 2025

Keywords:
assessment,
critical thinking,
test instruments,
independent curriculum,
Rasch model

Corresponding Author:
Nisa Us Sa'idah
Universitas Negeri Semarang
E-mail: nisasaidah89@students.unnes.ac.id

Abstract

This study aims to develop an assessment instrument for critical thinking skills in elementary school science lessons on the topic of Earth's changes, in the form of essay questions that are tested for validity and reliability. The measurement of critical thinking skills uses indicators based on Ennis' framework, which includes providing simple explanations, building basic skills, concluding, providing further explanations, and setting strategies and tactics. The subjects of this study were 28 students from SDN 02 Pekiringanalit, Pekalongan Regency. Data analysis techniques involved content validity testing using Aiken's index. The validity of the assessment instrument was tested using the Rasch Model with the Ministep program to determine validity, calculate reliability, and obtain Cronbach's alpha values. This research is developmental research that applies the 4D development model, simplified into three steps: define, design, and development. The results showed that the critical thinking assessment instrument in the form of essay test questions received an Aiken's content and language validity index of 0.96, which falls into the very valid category. The validity per item obtained an Aiken's index of 0.89, placing it in the high validity category. The item analysis using the Rasch Model yielded a person reliability coefficient of 0.69 and 0.72, which is in the sufficient category. The item reliability was 0.86 and 0.87, which is considered good, and the Cronbach's alpha was 0.77, also considered good.

ISSN 2775-2518

INTRODUCTION

The enhancement of superior human resources through education is a demand of the globalization era, addressing rapid changes in all fields. Twenty-first-century learning in elementary schools is essential for developing six core skills known as the 6Cs: character, critical thinking, creativity, citizenship, collaboration, and communication (Anggraeni et al., 2022). In the Merdeka Curriculum, critical reasoning is included in the dimensions of the Pancasila student profile (Ayu Gustianingrum & Murni, 2023). In this profile, students who can reason critically have the ability to process, relate, analyze, evaluate, and conclude information objectively, both qualitatively and quantitatively (Kurniawati et al., 2023).

Critical thinking is the ability to process information carefully to be applied wisely in decision-making and problem-solving (Heard et al., 2020). Students are equipped with critical thinking skills to solve problems in life (Rahmawati et al., 2023). Critical thinking ability is one of the essential skills to face the challenges of the 21st century(Defiyanti & Sumarni, 2019).

Science education in Indonesia is still below standard compared to other developing countries (Rusilowati et al., 2016). This is reflected in Indonesia's PISA scores in 2022, which show a decline, with reading scores reaching 359, mathematics scores reaching 366, and science scores reaching 383. (OECD, 2023). This indicates that students' ability to respond to questions requiring critical thinking skills in Indonesia is relatively low (Amelia & Magfirah, 2024). One of the main issues is the lack of teacher understanding of assessment types that can measure critical thinking skills. Teachers tend to create easy questions requiring short answers, thus neglecting students' critical thinking skills in the assessment process (Nugroho & Airlan, 2020). Additionally, students are not accustomed to questions that require them to think critically. This situation is influenced by the lack of practice in answering higher-order thinking questions that could enhance critical thinking skills, which should be provided by teachers.

Research conducted by (Lestari & Setyarsih, 2021) concluded that students' science literacy levels are in the sufficient category, while their critical thinking abilities are in the low category. This condition results in students being less active in exploring their thought processes. Therefore, it is recommended to develop science literacy and critical thinking skills through the use of essay test instruments that present concrete problems in daily life, such as those related to global warming.

Natural and Social Sciences (IPAS) in elementary schools play an essential role in shaping students' analytical and critical thinking patterns. IPAS helps students develop curiosity about occurring phenomena (Nuryani et al., 2023). Critical thinking is an important skill that every student must master, especially in understanding complex natural and social phenomena. Critical thinking involves a deep thinking process and the ability to draw conclusions based on evidence from everyday life issues (Lestari & Setyarsih, 2021). When provided with data or information, individuals can make accurate conclusions and identify possible contradictions, consistency, or deviations in the information (Sumarni & Kadarwati, 2020).

Measuring critical thinking skills in IPAS lessons, particularly in the topic of Earth's changes, remains a challenge. The lack of appropriate critical thinking assessment instruments tailored to the developmental characteristics of elementary school students often hampers evaluation.

The improvement of critical thinking skills is closely related to valid and reliable assessment instruments to measure students' learning progress and adjust teaching methods more effectively. Test scores can reflect students' learning outcomes in the classroom or comparisons to previously achieved performance (Sumintono, 2018). Assessing student progress in learning is a crucial action to evaluate their success (Safitri et al., 2023).

This study aims to develop valid and reliable assessment instruments to measure students' critical thinking skills in the IPAS subject on Earth's Changes at the elementary school level. Through the development of valid and reliable assessment instruments, it is expected to contribute to the improvement of the quality of IPAS learning in elementary schools. The developed assessment instrument is an essay test. The use of essay tests is chosen because this type of test encourages students to express their opinions according to their skills, thereby supporting the development of critical thinking skills (Sumarni et al., 2018).

METHOD

This study is a Research and Development (R&D) type of research. The development model used is the 4D model according to Thiagarajan (Thiagarajan et al., 1974) consisting of four stages: define, design, development, and dissemination. However, this study is limited to the stages of define, design, and development. The define stage involves needs analysis through preliminary studies and literature reviews. The design stage involves activities for designing the predetermined product. The development stage includes activities such as drafting question grids, creating questions, creating answer keys, developing assessment rubrics, conducting limited testing, and performing validation.

This study was conducted from January to May 2024 at SDN 02 Pekiringanalit. The subjects of this study were 28 sixth-grade students at SDN 02 Pekiringanalit, who served as the trial participants.

The research instruments included expert validation questionnaires and assessment instruments consisting of grids, questions, answer keys, and assessment rubrics.

Content validity of the assessment instruments was tested using Aiken's index (Retnawati, 2016). The formula for Aiken's index is:

$$V = \frac{\sum s}{n(c-1)}$$

Where:
V is the rater agreement index
s is the score assigned by each rater minus the lowest score in the category
n is the number of raters
c is the number of categories that raters can choose from
If the index is less than or equal to 0.4, the validity is considered low; 0.4-0.8 is considered moderate; and greater than 0.8 is considered highly valid (Retnawati, 2016).

Table 1. Criteria for Instrument Validity in Research

Aiken validity coefficient value	Category
$V < 0.4$	Less Valid
$0.4 \leq V \leq 0.8$	Valid
$V \geq 0.8$	Highly Valid

The content validity data used in this study were obtained from the validation results of the critical thinking assessment instrument for the IPAS subject on Earth's changes. This validation was conducted by four experts: one subject matter expert, two instrument experts, and one education practitioner.

The validity and reliability of the assessment instrument items were tested using the Ministep program and analyzed with the Rasch Model. The Rasch Model is an essential tool for teachers in developing test items and providing relevant information about student assessments for learning (Sumintono, 2018). The item validity analysis results are considered valid if at least two out of three criteria are met (Sari & Mahmudi, 2024).

Table 2. Criteria for Determining Cronbach's Alpha Values

Mark	Category
< 0.5	Poor
0.5 - 0.6	Bad
0.6 - 0.7	Fair
0.7 - 0.8	Good
> 0.8	Very good

(Sari & Mahmudi, 2024)

Table 3. Criteria for Determining Person Reliability and Item Reliability Values

Mark	Category
< 0.67	Weak
0.67 - 0.80	Fair
0.80 - 0.90	Good

0.90 - 0.94	Very Good
> 0.94	Excellent

(Sari & Mahmudi, 2024)

RESULTS AND DISCUSSION

The product generated is a critical thinking assessment instrument for the IPAS subject on Earth's changes in Elementary Schools. This assessment instrument is in the form of an essay test. It includes indicators of critical thinking skills according to Ennis, which are providing elementary clarification, building basic support, making inferences, providing advanced clarification, and managing strategies and tactics (Ennis, 1985). This instrument consists of grids, questions, and rubrics.

The results of the development of this assessment instrument can be accessed by clicking the link https://bit.ly/instrumenasesmenbumiberubah or scan the barcode in Figure 1.

Figure 1. The barcode of the developed instrument.

Content validity testing was conducted prior to validity testing using the Rasch model. This assessment instrument has been validated by 2 expert instrument lecturers, 1 expert subject matter lecturer, and 1 education practitioner. All experts stated that this instrument is highly valid for measuring critical thinking skills in the topic of Earth's changes. The instrument has been improved according to the guidance and suggestions from the experts. The following table shows the results of the content validity testing of the critical thinking skills assessment instrument.

Table 4. Results of Content Validity Testing

No	Assessment Aspects	Assessor Score				Aiken's V value	Information
		1	2	3	4		
A	Content validity						
	1. The material presented as problems is suitable for the school level or class level.	5	5	5	5	1	Highly Valid
	2. The problems can measure students' critical thinking abilities.	5	5	5	5	1	Highly Valid
	3. The clarity of the purpose of the presented problems.	5	4	5	5	0.94	Highly Valid
	4. The possibility of problems being resolved.	5	4	5	5	0.94	Highly Valid
B	Language Usage						
	1. The suitability of the language used in the problems with the rules of Indonesian language.	5	5	5	5	1	Highly Valid
	2. Sentences used in presenting the problems do not contain double meanings.	5	4	4	5	0.88	Highly Valid
	Average					**0.96**	**Highly Valid**

The analysis of the Aiken index yielded an average score of 0.96, falling within the range of values between 0.8-1, categorized as highly valid. Based on the expert validation results, the researcher made improvements to the product according to the experts' suggestions.

Furthermore, validation of each item was also conducted by experts on this instrument.

The results of the validity testing for each item can be seen in Table 5.

Table 5. Results of Content Validity Testing for Each Item.

Item No	Assessor Score				Aiken's Value	Description
	1	2	3	4		
1	4	3	4	4	0.92	High
2	4	3	4	4	0.92	High
3	4	3	4	4	0.92	High
4	4	3	4	4	0.92	High
5	4	3	4	4	0.92	High
6	4	3	4	4	0.92	High
7	4	3	4	4	0.92	High
8	4	3	4	4	0.92	High
9	3	3	4	4	0.83	High
10	4	3	3	4	0.83	High
average					0.89	High

Based on the validity test results using the Aiken index, the average score obtained was 0.89, falling within the range of values between 0.8-1, indicating high validity. Therefore, it can be concluded that all items are considered to have high validity. The trial results of the assessment instrument were then analyzed using Ministep software with Rasch Model analysis, as shown in Figure 2.

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item	
9	5	28	3.07	.67	.84	-.28	2.97	1.45	A .55	.59	88.5	88.3	S9	VALID
6	9	28	1.72	.52	1.28	1.10	2.16	1.54	B .48	.61	69.2	79.0	S6	VALID
10	18	28	-.44	.49	1.40	1.55	1.16	.49	C .44	.57	61.5	77.2	S10	VALID
2	16	28	.03	.48	.91	-.30	1.33	.90	D .60	.59	84.6	76.4	S2	VALID
3	23	28	-1.81	.57	1.18	.66	.77	.01	E .41	.45	73.1	83.0	S3	VALID
7	16	28	.03	.48	1.07	.38	.87	-.23	e .59	.59	69.2	76.4	S7	VALID
4	24	28	-2.17	.62	1.04	.22	.56	-.13	d .43	.41	80.8	85.8	S4	VALID
5	13	28	.73	.49	.98	-.02	.81	-.35	c .64	.61	73.1	76.3	S5	VALID
8	17	28	-.21	.49	.70	-1.29	.57	-1.12	b .70	.58	88.5	76.8	S8	VALID
1	20	28	-.95	.51	.49	-2.41	.34	-1.40	a .73	.53	96.2	78.7	S1	VALID
MEAN	16.1	28.0	.00	.53	.99	-.04	1.15	.12			78.5	79.8		
P.SD	5.6	.0	1.48	.06	.26	1.09	.78	.94			10.3	4.1		

Figure 2. Results of Rasch Model Item Validity Testing using Ministep Software

The criteria used to examine these items, including fit or misfit, can be conducted by analyzing the output from this Item fit order. At this stage, what needs to be considered for analysis are the values of Outfit Mean Square (MNSQ), Outfit ZStandard (ZSTD), and Point Measure Correlation (Pt Mean Corr). (Sari & Mahmudi, 2024). If an item meets three criteria or two of these criteria, then the item is considered valid (Sari & Mahmudi, 2024).

Data analysis was conducted using the Ministep application to analyze the Rasch model. Based on the analysis conducted, values for Cronbach's alpha, item reliability, and person reliability were found. The obtained values are shown in Figure 3.

```
     SUMMARY OF 28 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       5.8      10.0         .56     .95                                 |
|  SEM        .5         .0        .36     .06                                 |
| P.SD       2.6         .0       1.88     .30                                 |
| S.SD       2.6         .0       1.91     .31                                 |
| MAX.      10.0      10.0        4.55    1.93                                 |
| MIN.       1.0      10.0       -2.86     .75                                 |
|-----------------------------------------------------------------------------|
| REAL RMSE   1.05 TRUE SD    1.55  SEPARATION  1.48  Person RELIABILITY  .69 |
|MODEL RMSE    .99 TRUE SD    1.59  SEPARATION  1.60  Person RELIABILITY  .72 |
| S.E. OF Person MEAN = .36                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .77  SEM = 1.24
STANDARDIZED (50 ITEM) RELIABILITY = .93
```

Figure 3. Results of Rasch Model Item Validity Testing

The Cronbach's alpha value to measure reliability and describe the interaction between items and persons is 0.77, which falls into the good category. The person reliability value is 0.77, indicating a fairly good consistency in the respondents' answers.

Table 6. Criteria for Determining Cronbach's Alpha Values (Sari & Mahmudi, 2024).

Value	Category
< 0.5	Poor
0.5 – 0.6	Bad
0.6 – 0.7	Fair
0.7 – 0.8	Good
> 0.8	Very good

The analysis of item items with the Rasch Model yielded item reliability values of 0.86 and 0.87. This can be seen in Figure 4.

```
     SUMMARY OF 10 MEASURED (NON-EXTREME) Item
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      16.1      28.0         .00     .53       .99   -.04   1.15    .12 |
|  SEM        .9         .0        .49     .02       .09    .36    .26    .31 |
| P.SD       5.6         .0       1.48     .06       .26   1.09    .78    .94 |
| S.SD       5.9         .0       1.56     .07       .27   1.15    .82    .99 |
| MAX.      24.0      28.0        3.07     .67      1.40   1.55   2.97   1.54 |
| MIN.       5.0      28.0       -2.17     .48       .49  -2.41    .34  -1.40 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .56 TRUE SD    1.37  SEPARATION  2.45  Item   RELIABILITY  .86 |
|MODEL RMSE    .54 TRUE SD    1.38  SEPARATION  2.58  Item   RELIABILITY  .87 |
| S.E. OF Item MEAN = .49                                                     |
-------------------------------------------------------------------------------
```

Figure 4. Item Reliability

The analysis of test items using the Rasch Model produced item reliability values of 0.86 and 0.87, indicating a high level of consistency in item measurement. These values suggest that the items exhibit strong internal consistency, meaning that their difficulty levels and discrimination power remain stable across different respondent groups. High item reliability also reflects the adequacy of the test in measuring the intended construct with minimal measurement error.

Furthermore, the obtained reliability indices demonstrate that the Rasch Model effectively captures the underlying structure of the test items. A reliability value close to 1.00 implies that the test items are well-calibrated and consistently function as expected. This finding supports the robustness of the measurement instrument and its ability to provide valid and reliable data for further analysis.

Overall, the results indicate that the test instrument maintains a high level of reliability, ensuring consistency in item performance. The stability of item difficulty and respondent ability estimates further supports the suitability of the Rasch Model for item analysis. Future research may focus on refining item quality and exploring additional psychometric properties to enhance the overall effectiveness of the instrument.

CONCLUSION

The development of a critical thinking assessment instrument for IPAS lessons on Earth's Changing Material is a developmental research applying the Thiagarajan 4D development model, whose steps are simplified into define, design, and development. The define stage involves needs analysis through preliminary studies and literature reviews. The design stage entails designing the predetermined product. The development stage includes activities such as drafting question grids, creating questions, creating answer keys, developing assessment rubrics, conducting limited testing, and performing validation. The research findings indicate that the critical thinking assessment instrument in the form of essay test questions demonstrates high validity in terms of content and language, with an Aiken's V value of 0.96. The validity per item shows high validity with an Aiken's V index value of 0.89. The analysis of item questions using the Rasch model yielded person reliability coefficients of 0.69 and 0.72, which are categorized as sufficient. The item reliability values of 0.86 and 0.87 are categorized as good, and Cronbach's Alpha of 0.77 falls into the good category. The interpretation of the reliability analysis results indicates that the consistency of student responses is sufficient, and the quality of the item questions is good. The Cronbach alpha value falls within the good criteria, indicating that the developed instrument has good reliability coefficients. Critical thinking assessment refers to critical thinking skill indicators according to Ennis, which include providing simple explanations, building basic skills, drawing conclusions, providing further explanations, and managing strategies and tactics.

The developed assessment instrument does not yet cover all learning materials in IPAS lessons based on critical thinking indicators according to Ennis and other experts. The author suggests developing critical thinking skill assessment instruments for other materials to measure students' critical thinking abilities, thus allowing students to become accustomed to answering questions that enhance critical thinking skills.

REFERENCES

Amelia, A. R., Nasrah, N., & Magfirah, N. (2024). Pengaruh Model Pembelajaran Open-Ended Problem Terhadap Kemampuan Berpikir Kritis Siswa SD Pada Pembelajaran IPA. *Jurnal Riset dan Inovasi Pembelajaran*, *4*(1), 379-388.

Analysis of Students' Scientific Literacy Skills and The Relationship with Critical Thinking Skills on Global Warming Materials. In *Journal of Physics: Conference Series* (Vol. 1805, No. 1, p. 012040). IOP Publishing.

Anggraeni, P., Sunendar, D., Maftuh, B., Sopandi, W., & Puspita, R. D. (2022, March). Why 6 Cs? The urgency of learning at elementary school. In *4th International Conference on Educational Development and Quality Assurance (ICED-QA 2021)* (pp. 35-41). Atlantis Press.

Defiyanti, & Sumarni, W. (2019). Analisis Kemampuan Berpikir Kritis Setelah Penerapan Problem Based Learning Berbantuan Lembar Kerja Peserta Didik Bermuatan Etnosains. *Phenomenon: Jurnal Pendidikan MIPA*, *9*(2), 206-218.

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational leadership*, *43*(2), 44-48.

Heard, J., Scoular, C., Duckworth, D., Ramalingam, D., & Teo, I. (2020). Critical thinking: Skill development framework. https://research.acer.edu.au/ar_misc/

Kurniawati, W., Mardian Sungkari, F., Fitri Utami, A., Ria Adini, A., Puspitasari, L., Nurbiyanti, A., Pramudiyanti, H., Widiastuti, I., Septi Besdaningrum, D., Praptiwi , N., Vera Santi, E., Kholifah, E., & Marsanti, Y. (2023). *Science Learning in Elementary Schools*. Initiation of Berkarya Press.

Nugroho, AN, & Airlan, GS (2020). Development of an instrument for assessing critical thinking skills for grade 4 elementary school science learning. 3(3). https://doi.org/10.23887/jippg.v3i3

Rahmawati, H., Pujiastuti, P., & Cahyaningtyas, AP (2023). Categorization of Critical Thinking Ability of Fourth Grade Elementary School Students in Elementary Schools in Cluster II Kapanewon Playen, Gunung Kidul. *Journal of Education and Culture*, 8(1), 88–104. https://doi.org/10.24832/jpnk.v8i1.3338

Retnawati, H. (2016). Quantitative Analysis of Research Instruments. *Parama Publishing*. www.nuhamedika.gu.ma

Rusilowati, A., Kurniawati, L., Nugroho, SE, & Widiyatmoko, A. (2016). Developing an Instrument of Scientific Literacy Assessment on the Cycle Theme OPEN ACCESS. *In INTERNATIONAL JOURNAL OF ENVIRONMENTAL & SCIENCE EDUCATION* (Vol. 11, Issue 12).

Safitri, EM, Wahyuni, S., & Ahmad, N. (2023). Development of Critical Thinking Skills Assessment Instruments Using the Quizizz Application in Middle School Science Subjects. 14(1), 2086–7328.

Sari, EDK, & Mahmudi, I. (2024). Book: *Rasch Modeling Analysis in Educational Assessment*. PT Pena Persada Kerta Utama.

Sugih, S. N., Maula, L. H., & Nurmeta, I. K. (2023). Implementasi kurikulum merdeka dalam pembelajaran IPAS di sekolah dasar. *Jurnal Pendidikan Dasar Flobamorata*, *4*(2), 599-603.OECD. (2023). PISA 2022 Results (Volume I). OECD. https://doi.org/10.1787/53f23881-en

Sumarni, W., & Kadarwati, S. (2020). Ethno-stem project-based learning: Its impact on critical and creative thinking skills. *Indonesian Journal of Science Education,* 9(1), 11–21. https://doi.org/10.15294/jpii.v9i1.21754

Sumarni, W., Supardi, KI, & Widiarti, N. (2018). Development of assessment instruments to measure critical thinking skills. *IOP Conference Series: Materials Science and Engineering*, 349(1). https://doi.org/10.1088/1757-899X/349/1/012066

Sumintono, B. (2018, February). Rasch model measurements as tools in assesment for learning. In *Ist International Conference on Education Innovation (ICEI 2017)* (pp. 38-42). Atlantis Press.

Thiagarajan, S. (1974). *Instructional development for training teachers of exceptional children*: A sourcebook.