Jurnal Pendidikan Sains Indonesia
(Indonesian Journal of Science Education)
https://jurnal.usk.ac.id/jpsi/index

12(4), p.885-900, (2024)
e-ISSN: 2615-840X p-ISSN: 2338-4379
https://dx.doi.org/10.24815/jpsi.v12i4.39924

Development of Essential Thinking Test Instrument using the Presseisen Taxonomy on Ecosystem Material for Grade X High School Students

Afandi[1], Anisyah Yuniarti[1*], Azfa Fadhilah[1], Venny Karolina[2], Carla Cristiana Queiroz[3]

[1]Biology Education Departement, Faculty of Teacher Training and Education, Tanjungpura University, Pontianak, Indonesia.
[2]Social Science Department, Faculty of Teacher Training and Education, Tanjungpura University, Pontianak, Indonesia.
[3]Business and Administration Department, Economics and Business Faculty, Academia BAI, Luanda, Angola.

*Corresponding author's email: anisyah.yuniarti@fkip.untan.ac.id

Article History:
Received date: July 8 2024
Received in revised from: October 12 2024
Accepted date: October 15 2024
Available online: October 24 2024

Citation:
Afandi., Yuniarti, A., Fadhilah, A., Karolina, V., & Queiroz, C.C. 2024. Development of essential thinking test instrument using the presseisen taxonomy on ecosystem material for grade X High School Students. *Jurnal Pendidikan Sains Indonesia (Indonesian Journal of Science Education)*, 12(4):885-900.

Abstract. The test instrument is a measuring tool in the assessment process used to obtain data on student learning outcomes, both cognitive aspects and thinking skills. Test instruments to measure students' thinking skills are still rarely used. This study aims to develop an essential thinking test instrument using the Presseisen Taxonomy on valid and reliable ecosystem material. This study uses the R&D method with the 3D modification development model combined with the stages of test drafting which consists of 6 stages. Data was collected with interviews, tests, and test instrument validation sheets. The analysis included content validity, interrater reliability, and Rasch model analysis. The results showed that all items were valid based on content validity. Interrater reliability has a value of 0.783, which is a good category. All items are declared fit, but item number 7 contains bias, so it must be discarded. In order, Cronbach alpha, person reliability, and item reliability have the following values: 0.68 (enough), 0.50 (weak), and 0.99 (excellent). The item difficulty group consisted of two very difficult questions, four difficult questions, two easy questions, and two very easy questions. Thus, 9 out of 10 items developed are valid and reliable for use.
Keywords: Development, Essential Thinking, Presseisen Taxonomy, Rasch Model, Test Instrument.

Introduction

Assessment is an essential component of education. Assessment is a number of facts that explain the characteristics of someone or something (Erfianti et al., 2019). Proper assessment can affect student learning outcomes (Tan & Ong, 2020). Therefore, the implementation of the assessment must be carried out optimally in terms of techniques, methods, and quality of question items (Suyatna et al., 2020).

Assessment is part of the learning process that determines the success or failure of a learning process (Permatasari et al., 2019). One way to do this is by using test instruments. Test instruments can measure or determine student learning outcomes after or before the learning process (Erfianti et al., 2019). The use of test instruments aims to determine and identify students' level of understanding, as well as information that can be used to help improve the learning process.

Assessment of student learning outcomes is not only focused on cognitive aspects but can also measure skills that can support students' learning process, namely thinking skills. Thinking skills are one of the main objectives of education, and they aim to prepare students to understand concepts (Mubruroh & Suhandi, 2017). Thinking is an ongoing process that requires learners to solve problems using their minds or understanding, make good decisions, and take appropriate actions (Ramadhan et al., 2019). Thinking skills are part of analyzing and evaluating thoughts based on the work being done (Adri et al., 2022). Thinking skills are necessary in learning because they help build knowledge, solve problems, and formulate results (Chusni et al., 2019).

Thinking skills can be taught, learned and developed in the learning process with the assumption that learners can achieve goals and always evolving (Mubruroh & Suhandi, 2017). Efforts that can be made to assess students' success in developing thinking skills must be supported by measuring instruments that can measure these abilities (Mubruroh & Suhandi, 2017). One of the international standardized assessment models that can be used is a test instrument oriented toward higher order thinking skills (Setiawan et al., 2021). HOTS-level test instruments can be used to train students' to have higher-level thinking skills, which are the demands of the *merdeka* curriculum.

Based on the results of interviews with biology teachers at SMA Negeri 11, SMA Negeri 9, SMA Negeri 3, and SMA Negeri 5 Pontianak regarding the use of test instruments to measure students thinking skills, it is still not widely applied in schools. Information obtained from the interview results shows that the test questions used are still classified as low to medium cognitive level categories, namely C1-C3. This causes students to be unfamiliar with and challenged when working on HOTS-level questions, so students' thinking skills still need to improve. These problems are in line with research conducted by Baidlowi et al. (2019) that the questions given by teachers are only limited to measuring low-medium level thinking skills, namely at the C1 (remembering) and C2 (understanding) levels. Wikanta & Susilo (2022) stated that the low thinking ability of students is because the test questions used in cognitive assessment tend to test more aspects of memory, while test questions to train thinking skills are rarely used. Danczak et al. (2018) stated that the assessment is still traditional, so it has limitations in assessing students' thinking skills. Erfianti et al. (2019) stated that the form of instruments used for assessment is only limited to measuring students' knowledge with the indicators used at the lower order thinking skills level, namely C1, C2, and C3 so that students' understanding and higher-level thinking skills have not been appropriately measured.

Preparing test items must refer to a learning taxonomy (Feranda et al., 2021). There are various kinds of learning taxonomies, one of which is Bloom's taxonomy, often used to measure students' thinking skills. However, this study uses Presseisen taxonomy, which is still rarely used. This research is based on an umbrella research project to use learning taxonomies other than Bloom's taxonomy in the preparation of test instruments, whereas previous research conducted by Feranda et al. (2021) used the Marzano taxonomy and research by Camila et al. (2023) used the Stahl and Murphy taxonomy. Presseisen Taxonomy is one of the taxonomies that can be used to identify students' thinking skills from low to high cognitive levels, namely by using essential thinking skills indicators. Essential thinking, according to Presseisen taxonomy (1985), is a thinking process that includes five indicators, namely qualification (finding unique characteristics and providing qualifications), classification (making groupings), relationships (finding linkages), transformations (relating known to unknown characteristics creating new meanings), and causation (establishing cause and effect and interpretation).

Essential thinking is understanding and analyzing the core or essence of a problem through in-depth analysis so that it can be adequately resolved. Chusni et al. (2020) stated that essential thinking skills are thinking skills that students must master because the indicators of essential thinking are included in the basic skills needed in the thinking process

before reaching the next level of thinking. This statement is also reinforced by Presseisen (1985), which states that essential thinking skills are fundamental in the thinking process needed before going to a higher or more complex level of thinking.

Biology learning is one of the learning concepts that can be used to train thinking skills, reasoning, discovery, and connecting various related concepts (Rahmayumita & Hidayati, 2023). One of the concepts in learning biology is ecosystem material. Ecosystem material is quite complex and related to environmental problems that occur due to changes in the ecosystem. Through ecosystem material, students can deal directly with environmental problems and try to find solutions (Putri et al., 2023). Environmental problems require students to provide ideas or solutions to solving problems so that ecosystem material can be used to develop students' thinking skills (Fadillah, 2017). Previous research related to the development of test instruments on ecosystem material focused on problem-solving skills (Modok et al., 2021; Rahma et al., 2021), critical thinking skills (Putri et al., 2023; Wakhidah & Indana, 2021), and HOTS (Wardany et al., 2015). However, test instruments related to essential thinking skills using essential thinking indicators from the Presseisen Taxonomy have yet to be developed.

The novelty of this research lies in the urgency of developing instruments to improve students' essential thinking skills by using essential thinking indicators based on the Presseisen Taxonomy. Essential thinking indicators are suitable for developing test instruments on ecosystem material because essential thinking skills involve a fundamental and gradual thinking process that can be used to help find solutions to problems or issues and are suitable for helping students improve their thinking skills.

Based on the problems that have been described, researchers are interested in conducting research aimed at developing essential thinking test instruments using Presseisen's taxonomy on valid and reliable ecosystem material. This research is expected to be a reference for teachers in preparing test instruments, become an alternative in conducting the assessment process, and train and improve essential thinking skills through practice and application of students' knowledge.

Methods

This research is a R&D study using the 4D model by Thiagarajan et al. (1974), which was modified into a 3D model combined with the stages of test drafting according to Mardapi (2008), which consists of nine stages modified into six stages. The 3D model includes the define, design, and develop stages. The stages of test development include compiling test specifications, writing test questions, reviewing test questions, conducting test trials, analyzing test items, and improving tests.

The population of this study was all students of class X SMA in Senior High Schools in Pontianak City in the 2023/2024 school year, which amounted to 3868 people (based on the results of observations in each State Senior High School in Pontianak). The sampling technique used was the cluster sampling technique, namely sampling from a population based on the division of clusters or regions. The area was divided into five sub-districts in Pontianak City; each sub-district is represented by one school: SMA Negeri 3, SMA Negeri 5, SMA Negeri 8, SMA Negeri 9, and SMA Negeri 11 Pontianak. The determination of the sample size was calculated using the Slovin formula (Fauzy, 2019) as follows:

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

Remark:
n : number of samples used
N : total population
e^2 : magnitude of error in sampling (5%)

Based on the calculation, the sample used was 363 people divided into five schools, as shown in Table 1.

Table 1. Sampling Technique Using Cluster Sampling

Cluster	School	Sample Quantity
West Pontianak Subdistrict	SMA Negeri 11 Pontianak	73
East Pontianak Subistrict	SMA Negeri 9 Pontianak	72
Pontianak City Subdistrict	SMA Negeri 8 Pontianak	73
North Pontianak Subdistrict	SMA Negeri 5 Pontianak	72
South Pontianak Subdistrict	SMA Negeri 3 Pontianak	73
Southeast Pontianak Subdistrict	-	-
Total Sample		363

Data was collected with interviews, tests, and test instrument validation sheets. The instruments used in this study consisted of interview guidelines, test instruments, and test instrument validation guidelines. The interviews were structured using interview guidelines. The interviewees consisted of five biology teachers from the sample schools. The test instrument was prepared using essential thinking indicators according to Presseisen Taxonomy (1985), which consists of five indicators: qualification, classification, relationships, transformations, and causation. The test instrument consists of ten items on ecosystem material. The test instrument validation guideline sheet includes material, construct, and language aspects developed into eleven statement items.

The data analysis in this study consisted of content validity, interrater reliability, and Rasch model analysis. The validation process was carried out through two lecturers of Biology Education at from the Faculty of Teacher Training and Education, Tanjungpura University, and five Biology teachers. Validation uses a Likert scale of four criteria, as shown in Table 2.

Table 2. Statement Rating Scale

Criteria	Score
Strongly agree	4
Agree	3
Disagree	2
Strongly disagree	1

(Source: Sugiyono, 2022)

The data obtained was then analysed using Aiken's V formula (Aiken, 1985) as follows:

$$V = \frac{\sum s}{[n(c-1)]} \qquad (2)$$

Remark:

s : $r - l_o$

r : the score given by rater

l_o : the minimum validity score (1)

n : the number of raters

c : the maximum validity score (4)

The minimum validity value is determined using Aiken's V table. Question items are declared valid if they meet the minimum validity value with seven raters and four scales, which is 0.76 ($V \geq 0.76$).

Interrater reliability analysis was conducted using the IBM SPSS version 27 program based on the intraclass correlation coefficient (ICC), indicated by the average measure

value. The average measure value obtained is then grouped based on the categories shown in Table 3.

Table 3. Reliability Categories of ICC Test

ICC Value	Interpretation
›0.5	Poor reliability
0.5 ≤ICC≥ 0.75	Moderate reliability
0.75 ≤ICC≥ 0.90	Good reliability
›0.9	Excellent reliability

(Source: Koo & Li, 2015)

Rasch model analysis was conducted with the help of Winsteps version 3.73 application, which includes item fit analysis, reliability analysis (Cronbach alpha, person reliability, and item reliability), and item difficulty analysis (item measure).

The item suitability level is analyzed to determine valid (good) items and items that need to be revised. Items are valid if they meet one of the three criteria (Sumintono & Widhiarso, 2015). The criteria used to see the level of item fit in Rasch modeling are seen from the resulting values in outfit mean-square, outfit z-standard, and point measure correlation as shown in Table 4.

Table 4. Criteria of item fit level

Criteria	Logit Limit Values
Outfit Mean Square (MNSQ)	0.5 <MNSQ< 1.5
Outfit Z-Standard (ZSTD)	-2.0 <ZSTD< +2.0
Point Measure Correlation (Pt Measure Corr)	0.4 <Pt Measure Coor< 0.85

(Source: Bonee et al. in Sumintono & Widhiarso, 2015)

Reliability analysis is a test to determine the stability and consistency of the measurement results. The reliability value is indicated by the Cronbach alpha value, person reliability, and item reliability, which can be seen in Tables 5 and 6.

Table 5. Cronbach Alpha Criteria

Score	Criteria
› 0.8	Very good
0.7 – 0.8	Good
0.6 – 0.7	Enough
0.5 – 0.6	Bad
‹ 0.5	Very bad

(Source: Sumintono & Widhiarso, 2015)

Table 6. Item Reliability and Person Reliability Criteria

Score	Criteria
› 0.94	Special
0.91 – 0.94	Very good
0.81 – 0.90	Good
0.67 – 0.80	Enough
‹ 0.67	Weak

(Source: Sumintono & Widhiarso, 2015)

The level of difficulty indicates the level of difficulty of the item. The grouping of items is based on the standard deviation value obtained from the results of item measure analysis with categories of very difficult, difficult, easy, and very easy.

Results and Discussion

Define Stage

The define stage is carried out by conducting initial research and literature studies on the use of test instruments in schools, students' thinking skills, and the development of test instruments using essential thinking indicators according to Presseisen's Taxonomy to determine the needs of the test questions to be developed. The define stage includes initial end analysis, concept analysis, and analysis of learning objectives. Front-end analysis was conducted by collecting information through interviews with Biology teachers in grade X at SMA Negeri 3 Pontianak, SMA Negeri 5, SMA Negeri 8, SMA Negeri 9, and SMA Negeri 11 Pontianak to find out the problems faced in schools regarding the use of test instruments in the assessment process, students' thinking skills, and the urgency of developing test instruments. The information obtained was that the use of test instruments to measure thinking skills was still rarely used, and the test questions used by teachers were still classified as low-intermediate cognitive levels (C1-C3), which caused students' thinking skills to be low. In addition, there has been no development of test instruments that use essential thinking indicators, according to Presseisen's taxonomy. Concept analysis aims to identify the main concepts in ecosystem material, which consists of ecosystem components, forms of interaction, energy flow, biogeochemical cycles, and ecosystem changes (succession).

Learning objective analysis is an analysis carried out to determine learning objectives. This stage is carried out by reviewing the learning outcomes of the *Merdeka* curriculum to develop learning objectives. Furthermore, learning objectives are used to develop question indicators as the basis for preparing the developed test instruments.

Design Stage

The design stage is the product design stage, where the initial design is developed. It consists of compiling test specifications and writing test questions. Developing test specifications includes determining test objectives, compiling test grids, selecting test forms, and determining test length. The test in this study is a measuring tool to assess students' essential thinking skills on ecosystem material, as a teacher reference, and to train students' to improve essential thinking skills. The test grid is used as a guide when making test questions. The test grid contains identity, learning outcomes, learning objectives, essential thinking indicators, submaterial, question indicators, and question numbers. The test form developed in this study is a descriptive test form. The selection of the description test form is adjusted to the demands of the essential thinking indicator aspects and the purpose of test development. Descriptive tests are suitable for measuring complex learning outcomes and provide opportunities for students' to answer questions according to their way of thinking to train their thinking skills (Ropii & Fahrurrozi, 2017). The test length is the estimated time needed to complete the essential thinking test instrument, which is 90 minutes.

Writing test questions is a stage in which the question indicators are described into question sentences based on the grids that have been made. The question items developed were ten description test items on ecosystem material that referred to the essential thinking indicators according to Presseisen's Taxonomy, namely qualification, classification, relationship, transformation, and causation. The items are arranged based on problems related to everyday life sourced from news fragments, pictures, and discourse. The problems presented are linked to the essential thinking indicators that trigger students to use their thinking skills.

Develop Stage

The develop stage is for developing products designed to produce products. This stage includes reviewing test questions, conducting test trials, analyzing test items, and improving test questions.

Reviewing test questions is done by analyzing content validity. Validation is a process or activity to measure the validity level of products developed by raters (Junika et al., 2020). Validation aims to obtain rater approval of the feasibility of the product developed so that it can be used (Zakiyyatulmuna et al., 2022). Content validity analysis is carried out based on the assessment scores given by the raters. The score was then calculated using Aiken's V formula. The results of the content validity analysis can be seen in Table 7.

Table 7. Content Validity Analysis Results

Essential Thinking Indicator	Item No.	Aiken's V	Remark
Qualification	1	0.81	Valid
	5	0.81	Valid
Classification	2	0.91	Valid
	7	0.83	Valid
Relationship	3	0.83	Valid
	8	0.93	Valid
Transformation	6	0.87	Valid
	9	0.90	Valid
Causation	4	0.86	Valid
	10	0.88	Valid

Based on Table 7, the results of the content validity analysis on each item shows that the value obtained is greater than the minimum validity value (> 0.76), so each item developed is suitable for use to proceed to the next stage, namely, conducting a test of test questions.

In addition to content validity analysis, an interrater reliability analysis was also carried out at this stage. The results of the interrater reliability analysis obtained from the ICC output on the average measure value are 0.783. The value obtained is grouped based on the category that refers to Koo & Li (2015), which is the category of good reliability. In the ICC output results, reliability statistics output is also obtained, as seen in Cronbach's Alpha value. The instrument can be said to be reliable if the value of Cronbach's Alpha >0.60, meaning that if the value obtained >0.60, then the instrument is declared consistent and reliable, whereas if the value obtained <0.60, then the instrument is declared inconsistent and unreliable (Slamet & Wahyuningsih, 2022). Based on the analysis results, Cronbach's Alpha value is 0.775>0.60, so the essential thinking test instrument is declared reliable. After the essential thinking test instrument developed was analyzed for content validity and interrater reliability, then the test was carried out. Before being tested, the items were revised first. Item revisions are based on criticisms and suggestions given by the raters.

The test of test questions is carried out to see the suitability of the items (item fit), reliability, item difficulty (item measure), and detect bias in the items. The results of the test implementation can provide information that can be used to determine the standard of the items (Kirom & Hasyim, 2021), namely, knowing whether the items are suitable for use or still need to be corrected (Ulfah et al., 2020). The test questions that have been validated and revised are then tested on a research sample of 363 students.

The trial was conducted on ten description questions, with a processing time of 90 minutes. Students were asked to answer the test items on the answer sheet provided. Furthermore, students' answers were corrected in accordance with the rubric for scoring test questions. The scores obtained are then recapitulated into raw data, which will then

be analyzed with the Rasch model using Winsteps software version 3.73. One of the questions developed can be seen in Figure 1.

9. Perhatikan gambar dan penggalan kasus perubahan ekosistem di bawah ini!

Tumpukan Sampah Pasar Mawar Mengeluarkan Aroma Busuk di Tengah Kota Pontianak, Masyarakat: Baunya Tidak Seperti Mawar

Kelvin Novandi – 12 Februari 2023, 15:17 WIB

Pontianak -- Minggu (05/02/2023), Penumpukan sampah kembali terjadi di Tempat Penampungan Sementara (TPS) Pasar Mawar, Kota Pontianak. Penumpukan ini telah terjadi sejak hampir setiap pekan sehingga mengeluarkan aroma yang menyengat.

Menurut pantauan pewarta, sampah tersebut juga telah menumpuk hampir melebar ke ruas jalan Hos Cokroaminoto yang digunakan masyarakat Kota Pontianak menuju pasar maupun sebagai salah satu akses pusat kota.

Sampah tidak hanya mengeluarkan aroma yang mengganggu masyarakat sekitar namun juga menumpuk tinggi hingga 1,5 m hingga 2 m dan terus melebar ke ruas jalan. Selain itu, banyaknya sampah yang didominasi sampah plastik dan sampah organik seperti sayur dan buah yang membusuk menimbulkan aroma tak sedap sehingga beberapa pengendara yang melewati TPS tersebut harus menahan nafas mereka.

(Sumber: https://www.kompasiana.com/)

Kasus penumpukan sampah yang disajikan merupakan bentuk perubahan ekosistem. Buatlah perkiraan langkah yang tepat untuk mengatasi permasalahan pada kasus penumpukan sampah (minimal 4)!

Figure 1. An example of an item written in Indonesian

Item analysis is carried out to know the quality of the items (easy or difficult question categories), the effectiveness of the questions, and diagnostic information for students in distinguishing the level of students' abilities (students' understanding of the material that has been studied) (Fauziana & Wulansari, 2021). Item analysis is carried out with Rasch modeling to determine the suitability of the items (item fit), reliability, and items difficulty (item measure).

The Suitability of the Items (Item Fit)

Item fit in the Rasch model is also known as item validity. The item fit analysis in the Rasch model explains whether the test items function normally or not to make a measurement (Kirom & Hasyim, 2021), meaning that the test instrument can measure the variables under study correctly. If, based on the results of the analysis, there are items

that do not fit, it shows that students have misconceptions about the items (Sumintono & Widhiarso, 2015).

Question items can be categorized as fit if they meet one of the three criteria. However, if some items do not meet all three criteria, they have poor quality, so they need to be revised, replaced, or discarded (Sumintono & Widhiarso, 2015). The analysis results of the item fit output can be seen in Table 8.

Table 8. Items Fit Analysis Result

Item No.	*Outfit*		Pt. Measure Corr	Interpretation
	MNSQ	ZSTD		
1	1.10	1.4	0.43	Accepted
2	1.38	3.6 (*misfit*)	0.47	Accepted
3	0.78	-3.5 (*misfit*)	0.46	Accepted
4	1.06	0.9	0.42	Accepted
5	1.05	0.7	0.61	Accepted
6	1.00	0.1	0.39 (*misfit*)	Accepted
7	1.04	0.6	0.40 (*misfit*)	Accepted
8	1.02	0.3	0.44	Accepted
9	0.76	-3.0 (*misfit*)	0.52	Accepted
10	0.95	-0.6	0.43	Accepted

Based on Table 8, the test items are overall accepted, but some items do not meet one of the three criteria. The item number is retained because it only does not meet the criteria (misfit) on one of the criteria. Items with item codes E2, E3, and E9 fall into the misfit category because they have outfit ZSTD values outside the appropriate criteria (-2.0 <ZSTD< +2.0). Items with item codes E6 and E7 fall into the misfit category because they have an outfit value of Pt Measure Corr outside the appropriate criteria (0.4 <Pt Measure Corr< 0.85). These items were retained because they only did not meet the criteria (misfit) on one of the criteria.

According to research conducted by Tyas et al. (2020), four items did not meet the MNSQ criteria, and one item did not meet the Pt Measure Corr criteria. However, these items are within reasonable limits and do not need to be replaced or corrected. In research conducted by Lestari et al. (2023), one item does not meet the MNSQ and Pt Measure Corr criteria, but the item does not need to be replaced or changed because it still meets one of the three criteria used. Based on the results of research conducted by Tyas et al. (2020) and Lestari et al. (2023) are in line with the results of the research obtained and reinforced by the statement of Sumintono & Widhiarso (2015) that items included in the fit category must meet one of the three criteria used so that item numbers 2, 3, 6, 7, and 9 are still included in the fit category.

Question items can be considered valid if the question items do not contain bias (Ummah et al., 2022). Question items are said to be biased if they favor one individual with certain characteristics (Sumintono & Widhiarso, 2015). The Rasch model analysis used to detect bias is called differential item functioning (DIF) detection. The item contains bias if the probability value of the DIF output obtained is <5% (Sumintono & Widhiarso, 2015). In this study, whether or not the items contained bias was detected in gender characteristics. The results of the DIF item output analysis can be seen in Table 9.

Table 9. Items DIF Analysis Result

Item No.	Probability	Interpretation
1	0.0597	Accepted
2	1.00	Accepted
3	1.00	Accepted

4	1.00	Accepted
5	1.00	Accepted
6	0.2857	Accepted
7	0.0483	Rejected
8	1.00	Accepted
9	0.5666	Accepted
10	1.00	Accepted

Based on Table 9, the item detected to contain bias is item number 7, which has a probability value of 0.0483<0.05. Further information can be seen in the Figure 2.

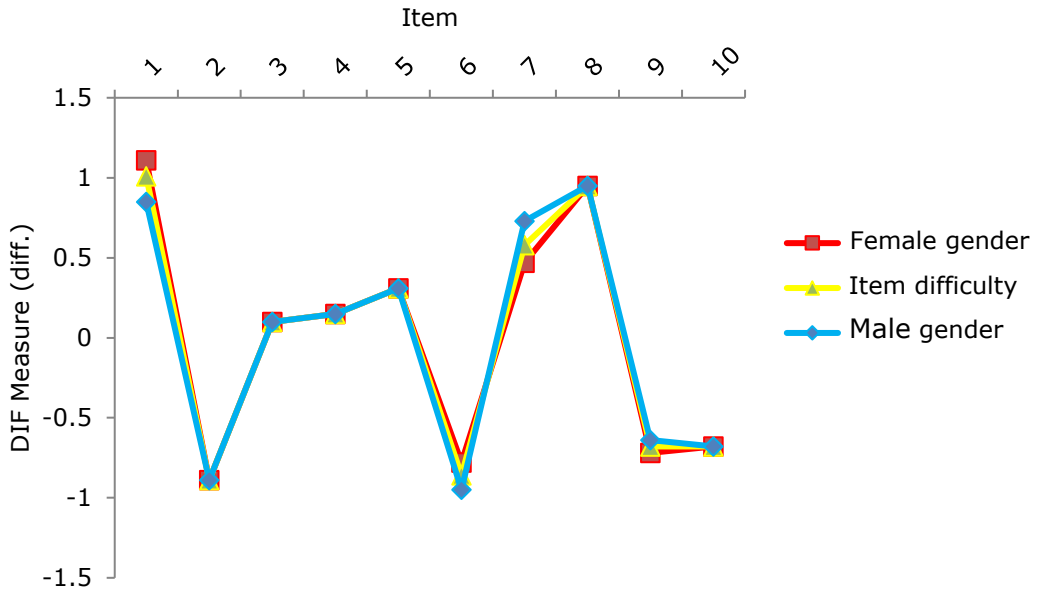

Figure 2. Graph of item bias based on gender characteristics.

Based on Figure 2, item number 7 has a biased tendency, where the L curve (male gender) and the P curve (female gender) have a long-distance DIF value. This shows that item number 7 is easily done by female students (below) compared to male students (above). Rusilowati (2018) states that the bias indicator can be seen from the partiality of the instrument to individuals with certain characteristics. For example, questions are easier for male students to answer than female students. This is also in line with research conducted by Aprilia et al. (2020) that indicates that items with curves approaching the lower limit are easy to do, while items with curves approaching the upper limit indicate that the items are difficult to do. Meanwhile, research conducted by Kirom & Hasyim (2021) found that several items contained bias; these questions were easier for male students to do than female students, so the questions contained bias. Thus, item number 7 was detected to contain bias and had to be discarded.

Reliability

Reliability is a test to determine the stability and consistency of the measurement results, meaning that the instrument used is reliable for measuring research variables where the measurement results are consistent when measuring two or more times against the same phenomena using the same measuring instrument (Hakim et al., 2021; Sugiono et al., 2020). Sumintono & Widhiarso (2015) suggest measuring reliability is indicated by

the Cronbach alpha, person reliability, and item reliability generated from the summary statistics output. The criteria of alpha Cronbach can be seen in Table 5, while the criteria of person and item reliability can be seen in Table 6. The analysis results from the summary statistics output can be seen in Table 10.

Table 10. Reliability Analysis Result

Output Summary Statistics	Score	Criteria
Cronbach alpha	0.68	Enough
Person reliability	0.50	Weak
Item reliability	0.99	Special

Based on Table 10, the Cronbach's alpha value obtained of 0.68 is in the range of 0.6-0.7 (enough), meaning that the interaction between learners (person) and items (item) as a whole is in a fairly good category. The person reliability value obtained is 0.50<0.67, meaning that the consistency of students' answers is in the weak category. The item reliability value obtained is 0.99>0.94, meaning that the quality of the items is excellent.

The person reliability value obtained is included in the weak category due to the diverse answers of students, which were seen from the number of research samples used and different sample locations. Suryanto et al. (2017) state that two possibilities cause low person reliability values. First, students' answers have a high enough diversity (variance), so the analysis results assess the possibility that students' understanding of test items is not uniform. That is, the possibility of the instrument is not easy to understand in general, so it is considered weak in terms of person reliability. Second, the answers lead to low uniformity, indicating the sample variety.

Based on this statement, the analysis results obtained from the summary statistics output lead to the second possibility: the person reliability value of 0.50 is categorized as weak, while the item reliability value of 0.99 is categorized as excellent. This is because the research sample used and the answers obtained from students have a high level of uniformity. The results of this study are supported by the results of research by Suryanto et al. (2017), where the analysis results of the person reliability obtained were categorized as weak (0.60), while the item reliability value was categorized as good (0.85), which means that there were no problems with the instruments used and the low person reliability value was caused by a relatively high and diverse research sample.

Item difficulty level (Item Measure)

The level of difficulty indicates the level of difficulty of the item. The items' difficulty level is seen from the logit measure value of the item measure output. Item grouping is based on the standard deviation value obtained (Sumintono & Widhiarso, 2015). The standard deviation value obtained is 0.70, so the grouping of item difficulty levels can be seen in Table 11. The analysis results of the item measure output can be seen in Table 12.

Table 11. Measure Value Categories

Score Category	Degree of difficulty
Logit >0.70	Very difficult
0.0 <*Logit*< (+0.70)	Difficult
(-0.70) <*Logit*< 0.0	Easy
Logit <-0.70	Very Easy

Table 12. Measure Items Analysis Results

Essential Thinking Indicator	Item No.	*Measure Logit*	Interpretation
Qualification	1	1.01	Very difficult
	5	0.31	Difficult
Classification	2	-0.89	Very easy
	7	0.58	Difficult
Relationship	3	0.10	Difficult
	8	0.95	Very difficult
Transformation	6	-0.86	Very easy
	9	-0.68	Easy
Causation	4	0.15	Difficult
	10	-0.68	Easy

Based on Table 10, it is known that two items (1 & 8) are included in the category of very difficult questions with a percentage of 20%, four items (3, 4, 5, & 7) are included in the category of difficult questions with a percentage of 40%, two items (9 & 10) are included in the category of easy questions with a percentage of 20%, and two items (2 & 6) are included in the category of very easy questions with a percentage of 20%. The level of difficulty of question items developed on the results of the item measure analysis shows that the question items have an inconsistent level of difficulty. When associated with essential thinking indicators, the difficulty level is in order from easy to difficult: qualification, classification, relationship, transformation, and causation. The data used to analyze the level of difficulty of the items is the number of scores obtained by students after the items are tested (raw data) so that the level of difficulty of the items depends on the answers given by the students. This aligns with research conducted by Feranda et al. (2022) that the distribution of the difficulty level of the items in the analysis results depends on the students' answers. The research results by Rahmat et al. (2020) also stated that even though the items have the same indicator level, each item will produce analysis results with different difficulty levels because students' answers influence them.

Improving the test is a stage to improve the essential thinking instrument tested at the test phase of the test questions and analyzed (Agustika, 2018). At this stage, sorting and determining items suitable for use based on the analysis results is carried out. Based on the results of the item validity test (item fit), it is known that all the items developed have met the valid criteria. However, based on the results of the DIF item test with gender characteristics, one item has a bias index, namely item number 7, so item number 7 must be discarded. The final product of the essential thinking test instrument that has been improved consists of 9 items out of 10 items.

Conclusion

The essential thinking test instrument using the Presseisen Taxonomy developed shows the results of content validity for each item included in the valid category because it has a value> 0.76. Interrater reliability has a value of 0.783 with a good reliability category. The reliability statistics analysis obtained has a Cronbach's Alpha value of 0.775> 0.60, so the essential thinking test instrument is declared reliable. The level of item fit shows that all items are included in the fit category because they meet the MNSQ, ZSTD, and Pt Measure Corr criteria. The bias index analysis results from the DIF item output show that question number 7 was detected to contain bias with a probability value of 0.0483 <0.05, so the item must be discarded. The summary statistics output obtained a Cronbach alpha value of 0.68 with a sufficient category, a person reliability value of 0.50 with a weak category, and an item reliability value of 0.99 with an excellent category. The level of

difficulty of the items obtained was two items in the category of very difficult questions, four items in the category of difficult questions, two items in the category of easy questions, and two items in the category of very easy questions. Based on the overall analysis, nine items are declared valid and reliable.

References

Adri, J. & Abdullah, A.S. 2022. Critical thinking skills in performance-based assessment: Instrument development and validation. *Journal of Technical Education and Training*, 14(1):90-99. https://doi.org/10.30880/jtet.2022.14.01.008

Agustika, G.N.S. 2018. Pengembangan konstruksi dan validasi tes konsep dasar matematika. *Journal of Education Technology*, 2(1):40-44. https://doi.org/10.23887/jet.v2i1.13805

Aiken, L.R. 1985. Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, 45(1):131-142. https://doi.org/10.1177/0013164485451012

Aprilia, N., Susilaningsih, E., Priatmoko, S., & Kasmui, K. 2020. Desain instrumen tes pemahaman konsep berbasis HOT dengan analisis model Rasch. *Chemistry in Education*, 9(2):1-8. https://journal.unnes.ac.id/sju/chemined/article/view/39068/17404

Baidlowi, M.H., Sunarmi, S., & Sulisetijono, S. 2019. Pengembangan instrumen soal *essay* tipe *higher order thinking skill* (HOTS) materi struktur jaringan dan fungsi organ pada tumbuhan kelas XI SMAN 1 Tumpang. *Jurnal Pendidikan Biologi,* 10(2):57-65. https://doi.org/10.17977/um052v10i2p57-65

Camila, C.G., Afandi., 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

Chusni, M.M., Saputro, S., & Rahardjo, S.B. 2020. The conceptual framework of designing a discovery learning modification model to empower students essential thinking skills. *Journal of Physics: Conference Series,* 1467(1):1-9. https://doi.org/10.1088/1742-6596/1467/1/012015

Danczak, S.M., Thompson, C.D., & Overton, T.L. 2020. Development and validation of an instrument to measure undergraduate chemistry students critical thinking skills. *Chemistry Education Research and Practice*, 21(1):62-78. https://doi.org/10.1039/C8RP00130H

Erfianti, L., Istiyono, E., & Kuswanto, H. 2019. Developing lup instrument test to measure higher order thinking skills (HOTS) Bloomian for senior high school students. *International Journal of Educational Research Review*, 4(3):320-329. https://doi.org/10.24331/ijere.573863

Fadillah, E.N. 2017. Pengembangan instrumen penilaian untuk mengukur keterampilan proses sains siswa SMA. *Didaktika Biologi: Jurnal Penelitian Pendidikan Biologi,* 1(2):123-134. https://doi.org/10.32502/dikbio.v1i2.770

Fauziana, A. & Wulansari, A.D. 2021. Analisis kualitas butir soal ulangan harian di Sekolah Dasar dengan model Rasch. *Jurnal Ibriez: Jurnal Kependidikan Dasar Islam Berbasis Sains*, 6(1):10-19. https://doi.org/10.21154/ibriez.v6i1.112

Fauzy, A. 2019. *Metode sampling*. Tangerang Selatan: Penerbit Universitas Terbuka.

Feranda, E., Ningsih, K., & Afandi. 2020. Penggunaan taksonomi Marzano dalam membuat soal HOTS: sebuah kajian literatur. *Prosiding Seminar Nasional Pendidikan,* 2(1):1053-1061.

Hakim, R.A., Mustika, I., & Yuliani, W. 2021. Validitas dan reliabilitas angket motivasi berprestasi. *FOKUS: Kajian Bimbingan dan Konseling dalam Pendidikan*, 4(4):263-268. https://doi.org/10.22460/fokus.v4i4.7249

Junika, N., Izzati, N., & Tambunan, L.R. 2020. Pengembangan soal statistika model PISA untuk melatih kemampuan literasi statistika siswa. *Mosharafa: Jurnal Pendidikan Matematika*, 9(3):499-510. https://doi.org/10.31980/mosharafa.v9i3.632

Kirom, A. & Hasyim, M. 2021. Analisis butir soal sebagai standarisasi mutu Sekolah Dasar pada mata pelajaran PAI dengan menggunakan pendekatan Rasch model di SD Ma'arif NU Kecamatan Pandaan Pasuruan. *Jurnal Al-Murabbi*, 6(2):92-98. https://jurnal.yudharta.ac.id/v2/index.php/pai/article/view/2631/1881

Koo, T.K. & Li, M.Y. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Lestari, T.D., Hamdu, G., & Saputra, E.R. 2023. Analisis soal literasi numerasi menggunakan pemodelan Rasch konteks pemanasan global berbasis ESD untuk Sekolah Dasar. *Pendas: Jurnal Ilmiah Pendidikan Dasar*, 8(2):2489-2503. https://doi.org/10.23969/jp.v8i2.9270

Mabruroh, F. & Suhandi, A. 2017. Construction of critical thinking skills test instrument related the concept on sound wave. *IOP Conf. Series: Journal of Physics: Conf. Series,* 812:1-6. https://doi.org/10.1088/1742-6596/812/1/012056

Mardapi, D. 2008. *Teknik penyusunan instrumen tes dan nontes*. Yogyakarta: MITRA CENDIKIA Press.

Modok, S.G., Budiretnani, D.A., & Nurmilawati, M. 2021. Pengembangan instrumen asesmen keterampilan pemecahan masalah berdasarkan Greenstein pada materi ekosistem*. Efektor,* 6(2):1-6. https://doi.org/10.29407/e

Permatasari, A.K., Istiyono, E., & Kuswanto, H. 2019. Developing assessment instrument to measure physics problem solving skills for mirror topic. *International Journal of Educational Research Review*, 4(3):358-366. https://doi.org/10.24331/ijere.573872

Presseisen, B.Z. 1985. *Thinking skills throughout the curriculum: a conceptual design.* Research for Better Schools, Philadelphia. National Inst. Of Education (ED), Washington, DC.

Putri, S.M., Arsih, F., Fadilah, M., & Anggriyani, R. 2023. Validitas instrumen soal tes keterampilan berpikir kritis siswa kelas X pada materi komponen ekosistem dan interaksinya. *Jurnal Pendidikan Tambusai*, 7(3):24253-24261. https://doi.org/10.31004/jptam.v7i3.10451

Rahma, R.A.N., Nurmilawati, M., Primandiri, P.R., & Sulistiono. 2021. Pengembangan instrumen asesmen keterampilan pemecahan masalah peserta didik SMA Negeri 1

Kediri pada materi ekosistem. *Jurnal Biologi dan Pembelajarannya (JB&P)*, 8(2):64-71. https://doi.org/10.29407/jbp.v8i2.16924

Rahmat, A.A., Hamdu, G., & Nur'aeni, E. 2020. Pengembangan soal tes tertulis berbasis STEM dengan pemodelan Rasch di Sekolah Dasar. *Metodik Didaktik: Jurnal Pendidikan Ke-SD-An*, 16(1):29-40. https://doi.org/10.17509/md.v16i1.25099

Rahmayumita, R. & Hidayati, N. 2023. Kurikulum merdeka: tantangan dan implementasinya pada pembelajaran biologi. *Biology and Education Journal,* 3(1):1-9. https://doi.org/10.25299/baej.2023.12758

Ramadhan, S., Mardapi, D., Prasetyo, Z.K., & Utomo, H.B. 2019. The development of an instrument to measure the higher order thinking skill in physics. *European Journal of Educational Research*, 8(3):743-751. https://doi.org/10.12973/eu-jer.8.3.743

Ropii, M. & Fahrurrozi, M. 2017. *Evaluasi hasil belajar*. Lombok Timur: Universitas Hamzanwadi Press.

Rusilowati, A. 2018. Asesmen literasi sains: analisis karakteristik instrumen dan kemampuan siswa menggunakan teori tes modern Rasch model. *Seminar Nasional Fisika Universitas Riau Ke-3 Universitas Riau, 2018 September*, hal. 1-15.

Setiawan, J., Sudrajat, A., & Kumalasari, D. 2021. Development of higher order thinking skill assessment instruments in learning indonesian history. *International Journal of Evaluation and Research in Education*, 10(2):545-552. https://doi.org/10.11591/ijere.v10i2.20796

Slamet, R. & Wahyuningsih, S. 2022. Validitas dan reliabilitas terhadap instrumen kepuasan kerja. *Aliansi: Jurnal Manajemen dan Bisnis*, 17(2):51-58. https://doi.org/10.46975/aliansi.v17i2.428

Sugiono, S., Noerdjanah, N., & Wahyu, A. 2020. Uji validitas dan reliabilitas alat ukur evaluasi postur SG. *Jurnal Keterapian Fisik*, 5(1):55-61. https://doi.org/10.37341/jkf.v5i1.167

Sugiyono. 2022. *Metode penelitian kuantitatif, kualitatif, dan R&D*. Bandung: Penerbit Alfabeta.

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

Suryanto, R., Indriyani, Y., & Sofyani, H. 2017. Determinan kemampuan auditor dalam mendeteksi kecurangan. *Jurnal Akuntansi dan Investasi,* 18(1):102-118. https://www. academia.edu/download/71650714/pdf_45.pdf

Suyatna, A., Viyanti, V., & Rosidin, U. 2020. Optimizing Computer-Based Hots Instruments: An Analysis of Test Items, Stimulus, and Quiz Setting Based on Physics Teachers' Perceptions. *Universal Journal of Educational Research*, 8(3D):97–105. https://doi.org/10.13189/ujer.2020.081714

Tan, B. & Ong, D. 2020. Pediatric to adult inflammatory bowel disease transition: the Asian experience. *Intestinal research*, 18(1):11-17. https://doi.org/10.5217/ir.2019.09144

Tyas, E.H., Hamdu, G., & Pranata, O.H. 2020. Analisis soal pilihan ganda dengan menggunakan pemodelan Rasch untuk mengukur kemampuan siswa dalam mengurutkan bilangan pecahan di Sekolah Dasar. *Pedadidaktika: Jurnal Ilmiah*

Pendidikan Guru Sekolah Dasar, 7(2):1-12. https://doi.org/10.17509/pedadidaktika.v7i2.24773

Ulfah, M., Djudin, T., & Oktavianty, E. 2020. Pengembangan tes kemampuan berpikir kritis peserta didik pada materi hukum newton di SMP. *Jurnal Pendidikan dan Pembelajaran Khatulistiwa (JPPK),* 9(1):1-13. http://dx.doi.org/10.26418/jppk.v9i1.38682

Ummah, K., Mardhiya, J., & Mulyanti, S. 2022. Pengembangan instrumen tes penguasaan konsep representasi kimia pada lima indikator asam basa dari alam: analisis dengan Rasch model. *Jurnal Tarbiyah*, 29(2):212-225. https://doi.org/10.30829/tar.v29i2.1706

Wakhidah, T.N. & Indana, S. 2021. Validitas dan reliabilitas tes elektronik (*e-test*) materi ekosistem untuk mengukur kemampuan berpikir kritis siswa kelas X SMA. *Berkala Ilmiah Pendidikan Biologi (BioEdu)*, 10(1):171-176. https://doi.org/10.26740/bioedu.v10n1.p171-176

Wardany, K., Sajidan., & Ramli, M. 2015. Penyusunan instrumen tes *higher order thinking skill* pada materi ekosistem SMA Kelas X. *Proceeding Biology Education Conference*, 12(1):538-543.

Wikanta, W. & Susilo, H. 2022. Higher order thinking skills achievement for biology education students in case-based biochemistry learning. *International Journal of Instruction*, 15(4):835-854.https://e-iji.net/ats/index.php/pub/article/view/290/388

Zakiyyatulmuna, N., Ningsih, K., & Wahyuni, E.S. 2022. Kelayakan multimedia interaktif berbasis *android* pada materi fungi kelas X SMA. *Bioilmi: Jurnal Pendidikan*, 8(2):147-158. https://doi.org/10.19109/bioilmi.v8i2.13802