Al-Ishlah: Jurnal Pendidikan
Vol.18, 2 (June, 2026), pp. 2553-2566
ISSN: 2087-9490 EISSN: 2597-940X, DOI: 10.35445/alishlah.v18i2.9839

Development of an Instrument for Assessing Scientific Process Skills in the Vertebrate Zoology Course for Trainee Biology Teachers

Muhamad Kurnia Sugandi[1,5], Saiful Ridlo[2], Dyah Rini Indriyanti[3], Sri Ngabekti[4]

[1] Universitas Negeri Semarang, Semarang, Indonesia; kurniasugandi@students.unnes.ac.id
[2] Universitas Negeri Semarang, Semarang, Indonesia; saiful_ridlo@mail.unnes.ac.id
[3] Universitas Negeri Semarang, Semarang, Indonesia; dyahrini@mail.unnes.ac.id
[4] Universitas Negeri Semarang, Semarang, Indonesia; sri.ngabekti@mail.unnes.ac.id
[5] Universitas Majalengka, Majalengka, Indonesia; kurniasugandi@unma.ac.id

ARTICLE INFO

Keywords:

science process skills;
instrument development;
Rasch model;
vertebrate zoology;
biology teacher education

Article history:

Received 2026-04-07
Revised 2026-05-21
Accepted 2026-06-29

ABSTRACT

Science process skills (SPS) are fundamental competencies for prospective biology teachers because they support scientific inquiry, critical thinking, and evidence-based problem-solving. However, contextualized and psychometrically validated assessment instruments for measuring SPS in Vertebrate Zoology remain limited. This study aimed to develop and validate an essay-based SPS assessment instrument for the Pisces superclass. This research employed a Research and Development (R&D) design using Thiagarajan's 4D model. The instrument was developed through the stages of define, design, and develop and was evaluated with 80 prospective biology teachers from two Indonesian universities. Content validity was established through expert judgment using Aiken's V, while construct validity, reliability, item difficulty, and item functioning were analyzed using the Rasch measurement model with Winsteps software. The findings indicate that the instrument demonstrated satisfactory psychometric properties. Most items achieved acceptable content validity (Aiken's V > 0.70), while Rasch analysis confirmed that 15 of the 17 essay items met the construct validity criteria. The instrument demonstrated adequate person reliability (0.67) and high item reliability (0.91), with item difficulty distributed across a broad range of student abilities and no evidence of item bias. The developed instrument is a valid and reliable tool for assessing science process skills in Vertebrate Zoology and offers a contextualized approach to authentic assessment in biology teacher education. Further validation involving larger and more diverse samples and additional biology topics is recommended to strengthen its generalizability and psychometric robustness.

Corresponding Author:
Saiful Ridlo
Universitas Negeri Semarang, Semarang, Indonesia; saiful_ridlo@mail.unnes.ac.id

1. INTRODUCTION

The challenges of science education among university students in the 21st century are multifaceted, particularly in developing critical thinking, problem-solving, and scientific process skills. Despite being in higher education, many students exhibit low proficiency in essential scientific processes, such as hypothesis formulation and data analysis, primarily because the curriculum emphasizes theoretical knowledge over practical inquiry and investigative experience (Yunita & Mandasari, 2025; Arifin et al., 2025). Recent curriculum reforms, such as those implemented at the University of Hong Kong, aim to address these issues by integrating science education with a focus on the scientific process and interdisciplinary relevance, thereby preparing students for diverse career paths (Kwok, 2018). Students at university often struggle to integrate scientific concepts into real-world problem-solving, primarily because of an educational focus on rote learning, which hampers their ability to apply the scientific method effectively. Research indicates that science process skills (including observing, classifying, predicting, and interpreting data are crucial for developing critical thinking and scientific literacy, particularly among prospective science and biology teachers (Anbiya et al., 2023; Maryanto et al., 2026).

According to the TIMSS (Trends in International Mathematics and Science Study) 2015 survey, Indonesia ranks low in science, placing 48th out of 56 countries. Indonesian students' science ability has an average score of 500, which is below international standards (Mullis et al., 2020). Indonesian students' abilities in science learning are very limited, characterized primarily by their ability to recognize basic facts but not to communicate effectively, link concepts, or apply complex ideas. Highlighting these challenges can help the audience feel empathetic and motivated to support educational improvements, recognizing the need for targeted interventions (Suliyanah et al., 2021; Kadir et al., 2025).

Efforts to address these challenges require the development of a science education system that encourages critical thinking and scientific process skills, enabling students to improve their ability to think, work, and adopt a scientific mindset, laying the foundation for problem-solving (Ugli, 2024). The learning process should strive to balance students' abilities across cognitive, affective, and psychomotor domains by incorporating practical activities, such as laboratory experiments (Kostøl et al., 2023; Cleaver et al., 2025). Furthermore, studies demonstrate that hands-on experimental methods significantly enhance psychomotor skills, as evidenced by substantial improvements in students' abilities to conduct and analyse experiments (Nurhayati, Sartika, & Salim, 2025). The implementation of practical activities will be more effective when supported by an efficient laboratory design that can develop students' science process skills (Oliveira et al., 2023; Zulfa & Malik, 2025). One skill that needs improvement in preparing future teachers for the 21st century is science process skills.

Science process skills are essential in science education because they empower educators to foster students' understanding of scientific concepts and critical thinking, making their teaching more impactful and meaningful (Azzahra et al., 2024). Recent research underscores the significant role of science process skills in enhancing students' science literacy and problem-solving abilities, crucial for navigating 21st-century challenges. Highlighting this importance can inspire educators and policymakers to recognize their vital role in fostering these skills for future success (Suja, 2020; Razali & Sanusi, 2020; Azzahra et al., 2024).

Assessment is a fundamental component of the educational process, serving multiple roles that enhance both teaching and learning outcomes. It provides a systematic approach to measuring students' understanding, skills, and attitudes, thereby enabling educators to identify learning gaps and tailor their instructional strategies accordingly (Zhou, 2023). Formative assessments, in particular, are crucial because they offer immediate feedback, fostering a supportive learning environment and motivating students to engage more deeply with the material. Good evaluation methods do more than measure grades. They also help teachers improve their teaching methods, which leads to better education. By using new assessment tools and regularly evaluating students, teachers can improve student growth and ensure everyone has a fair chance to succeed (Jannah & Widyanti, 2024). Assessment results can provide educators with feedback, enabling them to refine their teaching strategies, methods, and materials to better meet their students' needs. Assessment also helps students understand their strengths and weaknesses, thereby boosting their motivation and engagement in learning. Recent research indicates that effective assessment systems can enhance higher-order thinking skills, scientific literacy, and students' active engagement in learning (Moore et al., 2024).

Essay-based assessments are considered more appropriate for measuring scientific process skills as they allow for a more in-depth exploration of students' scientific thinking abilities than multiple-choice tests or performance-based assessments. Emphasizing the importance of conceptual understanding can help educators see the value of this approach in fostering deep learning. Scientific process skills encompass not only conceptual understanding but also the ability to observe, interpret data, formulate hypotheses, design investigations, analyze results, and communicate scientific findings. These abilities require higher-order thinking skills, which are more readily demonstrated through essay-style answers

Students who actively apply scientific process skills can solve everyday problems using the scientific method. This is one of the main goals of science education (Centinkaya & Ozyurek, 2019). Prospective science teachers exhibit low proficiency in science process skills, which are crucial for effective science education. Research indicates that while targeted interventions, such as feedback during laboratory activities, can improve performance, many students still demonstrate inadequate mastery of these skills (Nurita et al., 2023). Biology is one of the branches of science that can be directed towards the use of scientific process skills to solve life problems. Science process skills are considered important because they encourage students to describe objects and events, ask questions, formulate hypotheses, test hypotheses through observation and experimentation, and communicate ideas to others (Inayah et al., 2020; Farach et al., 2021). The high level of diversity and the use of Latin terminology make it difficult for students to understand material on vertebrates. In addition, vertebrate learning is closely related to animal observation and identification activities based on morphology, anatomy, and physiology. The availability of vertebrate animal specimens for identification is a challenge in learning biology, especially on the concept of vertebrates, especially after the pandemic (Fauziyah et al., 2020).

Research gap in the development of instruments for science process skills specifically designed for Vertebrate Zoology learning, particularly for the Pisces superclass, for trainee biology teachers. Previous research has generally emphasized the importance of science process skills in science learning, without developing contextual instruments suited to the complex characteristics of vertebrate material and grounded in scientific observation. Furthermore, the majority of studies have focused primarily on measuring cognitive aspects rather than comprehensively on science process skills (Apriyani & Aini, 2025). Research on the development of instruments for science process skills in the form of essay tests, analyzed using the Rasch model, in the field of biology remains limited. However, the Rasch model is considered capable of providing a more objective, accurate, and in-depth analysis of the validity, reliability, level of difficulty, and discriminatory power of the instrument compared to classical analysis approaches (Cinantya & Sugiharto, 2026). Research on instruments for science process skills in vertebrate zoology, specifically in the superclass Pisces, is limited. This study addresses this gap by developing an essay-based instrument and analyzing its characteristics using the Rasch model to create a valid, reliable, and appropriate assessment tool for vertebrate zoology learning.

The development of a science process skills instrument in the form of an essay test is expected to provide a clear picture of prospective biology teachers' science process skills. The research questions posed include: (1) What is the content validity of the science process skills instrument? (2) What is the construct validity of the science process skills test? (3) What is the reliability of the science process skills test? (4) What is the level of difficulty of the science process skills test items? Moreover, (5) What is the discriminatory power of the science process skills test?.

2. METHODS

This study is a research and development (R&D) project. The application of this method aims to produce a model for developing instructional materials, specifically a science process skills essay test instrument (Sugiyono, 2019), and to evaluate the effectiveness of the resulting product. The R&D model used is the instructional material development model developed by Thiagarajan, known as the 4D model, selected for its systematic approach to instructional design and proven effectiveness in similar contexts. This model encompasses four development stages: defining, designing, developing, and disseminating (Thiagarajan, 1974). This method and model were selected to create an assessment instrument that evaluates process skills within the Pisces superclass.

Data analysis was conducted descriptively to describe the characteristics of the science process skills test developed, including content validity, construct validity, reliability, item difficulty level, and item differentiation. The research sample consisted of 80 prospective biology teachers who had taken the vertebrate zoology course at Universitas Majalengka and UIN Siber Syekh Nurjati Cirebon, selected through purposive sampling to ensure relevance to the study objectives. Data analysis was conducted using the Rasch model to evaluate the instrument's quality, including item fit, reliability, difficulty, and discrimination. Validity criteria such as the Outfit Mean Square (MNSQ) within 0.5–1.5, the Outfit Z-Standard (ZSTD) within −2.0 to +2.0, and the Point Measure Correlation (Pt Measure Corr) within 0.40–0.85 were used to ensure item appropriateness. Reliability was assessed through test-retest and item reliability to confirm response consistency and item stability. Item difficulty was analyzed via logit values, and the instrument's discriminative ability was evaluated using separation indices. The analysis was conducted with Winsteps software, which provided an objective and comprehensive assessment of the instrument's psychometric properties (Hafizi et al., 2023; Sheptian, Arthur, & Wardhani, 2024).

Content validity was assessed through expert judgment by five validators: a subject-matter expert, a learning assessment expert, a language expert, a technology expert, and an education practitioner. The content validity analysis of the science process skills test instrument used Aiken's V as follows:

$$V = \frac{\sum S}{[n(C-1)]}$$

Notes:
S = R – lo
n = number of raters
lo = lowest rating
C = highest rating
R = ratings given by raters

Table 1. Interpretation of Validity Index

Aiken Index Scale Range	Description
$V > 0.84$	Very Valid
$0.68 < V \leq 0.84$	Valid
$0.52 < V \leq 0.68$	Moderately Valid
$0.36 < V \leq 0.52$	Less Valid
$V \leq 0.36$	Not Valid

Source: (Aiken, 1985)

3. FINDINGS AND DISCUSSION

3.1 Findings

3.1.1 Results of Development

Development results are the achievements obtained after completing the research procedure. The resulting product is a vertebrate zoology assessment instrument in the pisces superclass, designed to measure the science process skills of prospective biology teacher students. An explanation of the assessment instrument development results is as follows.

a. Define

During the define stage of the 4D development model, a needs analysis was conducted to identify learning issues and the need to develop teaching materials. The analysis showed that teaching in Vertebrate Zoology, particularly regarding the Superclass Pisces, remains focused on conceptual understanding and rote learning. Consequently, students' ability to engage in the scientific process has not yet been optimally assessed. Recognizing this gap can inspire educators and researchers to improve assessment methods that truly reflect students' scientific skills. Learning activities involving observation, organism identification, classification, data interpretation, and scientific problem-solving require comprehensive measurement of scientific process skills. However, the evaluation instruments currently in use are limited to measuring cognitive aspects. They cannot capture students' abilities to observe, classify, interpret, formulate hypotheses, plan investigations, apply concepts, and communicate scientific findings. This highlights the importance of developing assessment tools that measure scientific process skills, encouraging educators and researchers to enhance evaluation practices.

The assessment instrument consists of a question grid, a list of essay questions, and scoring guidelines. This instrument is used to evaluate the learning achievement of prospective biology teachers in the aspect of science process skills. The assessment was conducted via a written test comprising 17 essay items.

b. Design

At this stage, the assessment instrument was designed. The process involves making question grids, initial drafts of questions, and assessment criteria. Question preparation grids are made by considering the identification of aspects of science process skills, indicators of science process skills, and details of each question. The design is shown in Table 2.

Table 2. Lattice of Science Process Skills Instrument, Pisces Superclass

No	Aspects of Science Process Skills (SPS)	Indicator	Question Number	Score
1	Observe	Students can identify the main morphological characteristics of Osteichthyes and Chondrichthyes.	1, 2	0-4
2	Classifying	Students can classify fish based on observed characteristics into Osteichthyes or Chondrichthyes.	3,4	0-4
3	Summarise	Students can conclude the main differences between Osteichthyes and Chondrichthyes based on their physical characteristics and habitats.	5, 6	0-4
4	Communicating	Students can present observational data on the types of fish in the classes Osteichthyes and Chondrichthyes in the form of tables or graphs.	7, 8	0-4
5	Interpreting data	Students can interpret data on the distribution of Osteichthyes and Chondrichthyes across habitats and morphologies.	9, 10	0-4
6	Asking questions	Students can ask relevant, evidence-based questions about the physical adaptations of Osteichthyes and Chondrichthyes to their environments.	11, 12	0-4
7	Designing an experiment	Students can design simple experiments to test hypotheses about the physical adaptations of Osteichthyes and Chondrichthyes to their environment.	13	0-4
8	Conducting experiments	Students can carry out experimental procedures to test hypotheses about specific aspects of the biology of Osteichthyes and Chondrichthyes.	14, 15	0-4
9	Analysing data	Students are able to analyze experimental data to determine whether the data support or reject the hypothesis about the adaptation of Osteichthyes and Chondrichthyes.	16, 17	0-4

The science process skills test instrument employed an analytic scoring rubric with five assessment criteria (0, 1, 2, 3, and 4) to ensure objective and consistent evaluation of students' responses. A score of 0 was assigned when students provided no response or when the answer was irrelevant to the question, indicating an inability to demonstrate the targeted science process skill. A score of 1 reflected a very limited understanding of the concept or skill, where responses showed substantial misconceptions, incomplete reasoning, or minimal evidence of scientific thinking. A score of 2 indicated a basic level of understanding, demonstrating partial mastery of the concept or skill but still containing considerable inaccuracies, omissions, or weaknesses in scientific reasoning. A score of 3 was awarded when responses demonstrated good understanding and relatively appropriate application of science process skills, although several minor conceptual or procedural errors remained. The highest score, 4, was assigned to responses that demonstrated comprehensive conceptual understanding, accurate scientific reasoning, and appropriate application of science process skills, with few errors. This scoring system was designed to capture variations in students' scientific competence systematically while improving assessment reliability and reducing subjectivity in scoring.

c. Develop

The development stage is the process in which the previously planned assessment instrument is developed into a product ready for testing. The steps taken in this stage analyze the content validity, as validated by five validators: material experts, educational measurement and evaluation experts, linguists, educational technology experts, and educational practitioners. The average scores of 5 validators from Aiken's V index are presented in Table 3.

Table 3. Average Score of Aiken V Index with 5 Validators

No	Material Expert	Measurement Expert	Language Expert	Technology Expert	Practitioner Educator	Σ(s - l)	Aiken's V
1	4	3	5	4	4	15	0.75
2	4	3	4	3	5	14	0.70
3	5	4	4	3	4	15	0.75
4	4	3	4	4	5	15	0.75
5	5	5	4	4	4	17	0.85
6	4	4	5	4	4	16	0.80
7	4	4	4	4	4	15	0.75
8	4	4	5	4	5	17	0.85
9	5	3	4	3	4	14	0.70
10	4	4	5	3	4	15	0.75
11	4	3	4	4	4	14	0.70
12	5	4	3	4	4	15	0.75
13	4	5	4	4	3	15	0.75
14	4	4	4	4	5	16	0.80
15	5	4	5	4	4	17	0.85
16	5	3	4	3	5	15	0.75
17	4	4	5	4	4	16	0.80

Aiken's V is an index used to measure the content validity of an instrument, particularly in the context of test or questionnaire development. Content validity refers to the extent to which the items in the instrument representatively cover the domain of the construct being measured (Aiken, 1985). Most items demonstrate good to very good content validity. The Aiken's V values for most items are at or above 0.75, indicating that experts have given positive evaluations of these items. Items 2, 9, and 11 have an Aiken's V value of 0.70, indicating valid content validity with good criteria. Items 1, 3, 4, 7, 10, 12, 13, and 16 have an Aiken's V value of 0.75, indicating valid content validity with good criteria. Items 5, 6, 8, 14, 15, and 17 have an Aiken's V value of 0.80–0.85, indicating very high content validity with excellent criteria.

3.1.2 Construct Validity

a. Validity test

The validity test determines whether a measuring instrument is valid. The validity of the test instrument, in the form of essay questions on aspects of science process skills, in the Pisces superclass was analyzed using the Rasch Model with the Winstep application. A summary of the validity test results using the Rasch Model assisted by the Winstep application can be seen in table 4.

Table 4. Science Process Skills Test Validity Test Results

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	
5	294	80	-.84	.22	1.27	1.30	1.91	3.00	A-.04	.25	64.6	69.7	P5	Invalid
11	278	80	-.23	.18	1.25	1.43	1.31	1.48	B .17	.31	44.3	53.9	P11	Valid
12	250	80	.52	.15	1.30	1.89	1.22	1.34	C .46	.39	40.5	49.0	P12	Valid
13	224	80	1.08	.14	1.20	1.37	1.17	1.11	D .55	.44	38.0	46.5	P13	Valid
15	250	80	.52	.15	1.16	1.04	1.15	.93	E .41	.39	38.0	49.0	P15	Valid
3	273	80	-.08	.17	.99	-.02	1.15	.81	F .19	.33	44.3	51.7	P3	Valid
7	243	80	.68	.15	1.10	.71	1.05	.36	G .44	.40	30.4	48.5	P7	Valid
4	280	80	-.30	.18	.84	-.93	.99	.00	H .04	.31	38.0	55.9	P4	Valid
14	294	80	-.84	.22	.98	-.03	.85	-.53	I .32	.25	69.6	69.7	P14	Valid
16	260	80	.28	.16	.96	-.18	.89	-.64	h .55	.36	43.0	49.0	P16	Valid
9	268	80	.07	.17	.94	-.33	.90	-.50	g .33	.34	43.0	50.3	P9	Valid
1	305	80	-1.49	.27	.92	-.22	.89	-.23	f .15	.20	79.7	82.0	P1	Invalid
8	270	80	.01	.17	.85	-.94	.79	-1.21	e .34	.34	51.9	51.7	P8	Valid
2	279	80	-.26	.18	.82	-1.08	.80	-1.04	d .33	.31	57.0	55.7	P2	Valid
10	249	80	.55	.15	.80	-1.43	.82	-1.14	c .43	.39	54.4	49.0	P10	Valid
17	260	80	.28	.16	.59	-3.11	.81	-1.12	b .37	.36	70.9	49.0	P17	Valid
6	269	80	.04	.17	.50	-3.78	.56	-2.86	a .43	.34	68.4	50.3	P6	Valid
MEAN	267.4	80.0	.00	.18	.97	-.25	1.01	-.01			51.5	54.8		
P.SD	20.0	.0	.62	.03	.22	1.51	.29	1.33			14.1	9.5		

Based on the results of the Rasch model analysis in the table, items 1 (P1) and 5 (P5) were deemed invalid because they did not meet the item fit criteria, specifically regarding the Mean Squared Error of Prediction (MNSQ), Z-Standard Error of Prediction (ZSTD), and Point Measure Correlation (Pt Measure Corr). Recognizing these issues emphasizes the importance of thorough validation to ensure assessment quality.

Item number 5 (P5) was deemed invalid because it had an Outfit MNSQ value of 1.91, which falls outside the ideal range of 0.5–1.5, and an Outfit ZSTD value of 3.00, which exceeds the acceptable limit of −2.0 to +2.0. Additionally, the Point Measure Correlation value of −0.04 indicates a negative correlation between the item and the science process skills construct. Addressing such issues highlights the importance of expert review to maintain assessment validity.

Meanwhile, item number 1 (P1) has a Point Measure Correlation value of 0.15, which is below the ideal range (0.40–0.85). Although the Infit MNSQ (0.92) and Outfit MNSQ (0.89) values are still within an acceptable range, the low correlation values indicate that the item is less capable of distinguishing between students with high and low abilities. This underscores the importance of precise item analysis to ensure measurement robustness.

Regarding follow-up, invalid test items are generally not immediately removed, especially in instrument development research. A more appropriate step is to revise the test items based on the Rasch and content analyses. This focus on revision can foster a sense of purpose and dedication, emphasizing the importance of continuous improvement (Rozali et al., 2022; Ubaidillah et al., 2022). Revisions can be made to the wording of the item, the clarity of the question's context, the alignment with science process skill indicators, or the item's level of complexity. After revision, the items need to be retested in the next pilot test phase to ensure improved instrument quality.

Thus, in this study, items 1 and 5 are recommended for revision and retesting rather than immediate deletion, as revision can improve their alignment with the science process skills construct being measured. Recognizing this effort can motivate researchers by highlighting their role in enhancing instrument validity.

b. Reliability Test

Reliability is an index that shows the extent to which a measuring device can be trusted or relied upon. The reliability test is used to determine the consistency of the measuring instrument, namely, whether the tool produces the same results across repeated measurements (Windi, 2011). High and low reliability are empirically indicated by a number called the reliability coefficient. The higher the correlation coefficient between the measurement results of two parallel tests, the higher the consistency between the two, and the measuring instrument is called reliable (Purnamasari, 2016). The results of the reliability test using the Rasch Model, assisted by the Winstep application, are shown in Table 5.

Table 5. Reliability Test Results of Science Process Skills Test

CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .67 SEM = 2.81
STANDARDIZED (50 ITEM) RELIABILITY = .87

SUMMARY OF 80 MEASURED (NON-EXTREME) Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	267.4	80.0	.00	.18	.97	-.25	1.01	-.01
SEM	5.0	.0	.15	.01	.06	.38	.07	.33
P.SD	20.0	.0	.62	.03	.22	1.51	.29	1.33
S.SD	20.6	.0	.64	.03	.23	1.56	.30	1.37
MAX.	305.0	80.0	1.08	.27	1.30	1.89	1.91	3.00
MIN.	224.0	80.0	-1.49	.14	.50	-3.78	.56	-2.86

REAL RMSE .19 TRUE SD .59 SEPARATION 3.19 Item RELIABILITY .91
MODEL RMSE .18 TRUE SD .59 SEPARATION 3.31 Item RELIABILITY .92
S.E. OF Item MEAN = .15

Based on the Rasch Model analysis, the Person Raw Score Reliability (test reliability) was 0.67. This figure helps educators and researchers understand the assessment's current consistency level, which is adequate but highlights room for improvement to ensure more reliable measurements of respondents' abilities.

However, the reliability value of 0.67 also indicates that participants' responses to the instrument are not yet consistent. Variation in student responses remains quite significant, so the instrument's ability to distinguish levels of science process skills among respondents is not yet fully robust. This may be influenced by several factors, such as items that are not yet fully aligned with the construct being measured, poorly suited items, or heterogeneity in participants' abilities.

Nevertheless, the instrument remains useful as a preliminary assessment tool. These results should motivate stakeholders to collaborate on refining the instrument by revising invalid or poorly aligned items, thereby enhancing test reliability in future administrations. This finding is supported by an item reliability coefficient of 0.91, indicating that the quality and difficulty of the test items are already in the "very good" range; therefore, any improvements should focus on improving the consistency of participants' responses to the instrument.

3.1.3 Problem Difficulty Level

The stage of testing the difficulty of the instrument aims to identify the level of difficulty of each question, whether it is easy, moderate, or difficult. A question that is considered good is one that has a balanced level of difficulty, not too easy and not too difficult. The results of the difficulty analysis of the questions are presented in the following table.

Table 6. Test Results of the Level of Difficulty of the Science Process Skills Test

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	
13	224	80	1.08	.14	1.20	1.37	1.17	1.11	.55	.44	38.0	46.5	P13	Very difficult
7	243	80	.68	.15	1.10	.71	1.05	.36	.44	.40	30.4	48.5	P7	Difficult
10	249	80	.55	.15	.80	-1.43	.82	-1.14	.43	.39	54.4	49.0	P10	Difficult
12	250	80	.52	.15	1.30	1.89	1.22	1.34	.46	.39	40.5	49.0	P12	Difficult
15	250	80	.52	.15	1.16	1.04	1.15	.93	.41	.39	38.0	49.0	P15	Difficult
16	260	80	.28	.16	.96	-.18	.89	-.64	.55	.36	43.0	49.0	P16	Medium
17	260	80	.28	.16	.59	-3.11	.81	-1.12	.37	.36	70.9	49.0	P17	Medium
9	268	80	.07	.17	.94	-.33	.90	-.50	.33	.34	43.0	50.3	P9	Easy
6	269	80	.04	.17	.50	-3.78	.56	-2.86	.43	.34	68.4	50.3	P6	Easy
8	270	80	.01	.17	.85	-.94	.79	-1.21	.34	.34	51.9	51.7	P8	Easy
3	273	80	-.08	.17	.99	-.02	1.15	.81	.19	.33	44.3	51.7	P3	Very easy
11	278	80	-.23	.18	1.25	1.43	1.31	1.48	.17	.31	44.3	53.9	P11	Very easy
2	279	80	.26	.18	.82	-1.08	.80	-1.04	.33	.31	57.0	55.7	P2	Medium
4	280	80	.30	.18	.84	-.93	.99	.00	.04	.31	38.0	55.9	P4	medium
5	294	80	.84	.22	1.27	1.30	1.91	3.00	-.04	.25	64.6	69.7	P5	Difficult
14	294	80	-.84	.22	.98	-.03	.85	-.53	.32	.25	69.6	69.7	P14	Very easy
1	305	80	-1.49	.27	.92	-.22	.89	-.23	.15	.20	79.7	82.0	P1	Very easy
MEAN	267.4	80.0	.00	.18	.97	-.25	1.01	-.01			51.5	54.8		
P.SD	20.0	.0	.62	.03	.22	1.51	.29	1.33			14.1	9.5		

Based on the Rasch model analysis, the items spanned difficulty levels from very difficult to very easy, demonstrating the instrument's capacity to assess students' science process skills across a broad ability spectrum. Item P13, with the highest difficulty (1.08 logits), is classified as very difficult, while P1, with the lowest difficulty (−1.49 logits), falls into the very easy category. This distribution indicates that the instrument has a sufficiently representative range of items to differentiate students' abilities. Nonetheless, the presence of some very easy or very difficult items may limit the instrument's sensitivity in accurately measuring abilities throughout the entire range.

This study also has limitations, namely the limited number of respondents and the fact that the instrument was developed only for the Pisces Superclass material; therefore, further testing on a broader sample and other Vertebrate Zoology materials is necessary to strengthen the findings. Additionally, several items with extreme difficulty levels require evaluation and refinement to optimize the instrument's measurement of science process skills, which is a natural step toward enhancing its validity (Komisia et al., 2025) (Zahra et al., 2025).

3.1.4 Differentiability Test

Differentiability test of an item refers to its ability to separate high-ability students from low-ability students. Evaluation of the differentiability test for each item is based on the item's DIF value or tendency. The results of the item score tendency test can be found in Table 7 below.

Table 7. Test of Differentiability of Science Process Skills

Person CLASSES	SUMMARY DIF CHI-SQUARED	D.F.	PROB.	BETWEEN-CLASS/GROUP UNWTD MNSQ	ZSTD	Item Number	Name
9	4.9007	8	.7681	.8608	-.13	1	P1
9	5.1303	8	.7435	.8105	-.24	2	P2
9	8.0881	8	.4248	2.7835	2.61	3	P3
9	14.0245	8	.0811	2.3779	2.18	4	P4
9	8.9745	8	.3443	1.5560	1.12	5	P5
9	4.4247	8	.8169	.8271	-.20	6	P6
9	14.4428	8	.0709	2.8365	2.66	7	P7
9	7.3197	8	.5025	1.1832	.51	8	P8
9	4.4695	8	.8124	.6839	-.55	9	P9
9	2.4552	8	.9638	.4060	-1.39	10	P10
9	6.9095	8	.5463	1.1596	.47	11	P11
9	8.5157	8	.3847	1.4290	.92	12	P12
9	8.8755	8	.3528	1.4482	.96	13	P13
9	10.5070	8	.2311	2.4803	2.29	14	P14
9	7.2717	8	.5075	1.3249	.76	15	P15
9	6.9648	8	.5404	1.1526	.46	16	P16
9	.7326	8	.9994	.1487	-2.65	17	P17

Based on Table 7, it can generally be stated that the science process skills test does not contain biased items, as all item probability values are greater than 0.05. Questions are said to be unbiased if they do not give certain individuals an advantage. Overall, based on the validity, reliability, question difficulty, and question differentiation tests, the Pisces Superclass Science Process Skills test instrument was declared valid and reliable. The developed instrument for science process skills can effectively measure prospective biology teachers' science process skills. This instrument yielded 15 valid questions out of 17.

3.2 Discussion

The present study developed and validated an essay-based instrument to assess science process skills (SPS) among prospective biology teachers in the Vertebrate Zoology course, specifically within the Pisces superclass. The findings demonstrate that the instrument possesses satisfactory psychometric properties, evidenced by acceptable content validity, adequate construct validity, and high item reliability. These results suggest that the instrument has the potential to serve as an authentic assessment tool for measuring students' ability to engage in scientific inquiry beyond simple conceptual recall. Unlike conventional cognitive assessments, the developed instrument captures essential dimensions of science process skills, including observation, classification, data interpretation, hypothesis formulation, experimental design, and scientific communication, which are fundamental competencies for future biology teachers (Azzahra et al., 2024; Moore et al., 2024). The findings therefore support the growing consensus that science assessment should emphasize inquiry competence and higher-order thinking rather than factual knowledge alone (Suja, 2020).

The content validation results further indicate that expert reviewers consistently judged the instrument to be relevant and representative of the targeted science process skills. Most items achieved Aiken's V values exceeding the recommended threshold, reflecting strong agreement among experts regarding content relevance and instructional appropriateness. This finding reinforces Aiken's (1985) argument that systematic expert judgment constitutes an essential step in establishing content validity during educational instrument development. The involvement of experts from multiple disciplinary backgrounds—including biology content, educational assessment, language, educational technology, and educational practice—also strengthened the comprehensiveness of the validation process. Such multidisciplinary validation is particularly important for essay-based assessments because item quality depends not only on scientific accuracy but also on linguistic clarity, construct alignment, and instructional feasibility (Hafizi et al., 2023). Consequently, the study demonstrates that combining expert judgment with empirical validation provides a stronger foundation for developing educational assessment instruments.

The Rasch analysis provides additional evidence regarding the instrument's construct validity while simultaneously identifying aspects requiring refinement. Fifteen of the seventeen essay items satisfied the recommended fit statistics, indicating that most items functioned consistently in measuring the intended science process skills construct. However, two items exhibited unacceptable fit indices and therefore require revision before wider implementation. Rather than representing a weakness, this finding illustrates the value of Rasch measurement in identifying problematic items that may not adequately represent the intended construct. Rasch analysis offers advantages over classical test theory because it evaluates individual item functioning, estimates item difficulty independently of respondent ability, and provides objective evidence regarding measurement quality (Bond et al., 2021; Boone et al., 2014). The identification of misfitting items enables researchers to improve item wording, contextual relevance, or cognitive demand, thereby strengthening the overall validity of the instrument. Accordingly, the present study demonstrates that psychometric evaluation should be viewed as an iterative process rather than a one-time validation procedure.

An interesting finding concerns the difference between item reliability and person reliability. The instrument achieved very high item reliability, indicating that the hierarchy of item difficulty was stable and consistently estimated across participants. In contrast, the person reliability coefficient remained in the moderate category, suggesting that the instrument was less effective in differentiating respondents according to their science process skills. This discrepancy may reflect relatively homogeneous participant abilities, limited sample variability, or the need for additional items representing intermediate levels of difficulty. Similar interpretations have been reported in Rasch measurement studies, where high item reliability accompanied by moderate person reliability often indicates that further refinement of item distribution or broader participant sampling is required to improve measurement precision (Bond et al., 2021; Boone et al., 2014). Therefore, while the developed instrument demonstrates promising psychometric quality, future validation involving more heterogeneous populations is necessary to enhance its discriminatory capacity.

The analysis of item difficulty also provides meaningful insights into the nature of science process skills among prospective biology teachers. The instrument contained items ranging from very easy to very difficult, allowing assessment across a broad spectrum of student abilities. Notably, items assessing experimental design were among the most difficult, whereas items requiring simple observation were comparatively easier. This pattern is theoretically consistent with science education literature, which positions experimental design, hypothesis formulation, and data analysis as integrated higher-order inquiry skills requiring complex scientific reasoning (Anbiya et al., 2023; Inayah et al., 2020). These findings suggest that although prospective biology teachers generally demonstrate competence in basic observational skills, they continue to experience difficulties when engaging in more sophisticated scientific reasoning tasks. Consequently, biology instruction should provide greater opportunities for inquiry-based laboratory activities and authentic scientific investigations that strengthen advanced science process skills.

The absence of statistically significant item bias further supports the fairness of the developed assessment instrument. Differential Item Functioning (DIF) analysis indicated that all items functioned consistently across respondents, suggesting that item performance was determined primarily by students' science process skills rather than by systematic bias. Fairness represents an essential characteristic of educational assessment because valid instruments should measure the intended construct without favoring particular groups of examinees (Hafizi et al., 2023). The present findings therefore indicate that the instrument possesses not only acceptable validity and reliability but also equitable measurement characteristics suitable for educational assessment within the studied context.

Overall, this study contributes to science education by providing a context-specific, essay-based assessment instrument that integrates expert validation with Rasch measurement to evaluate science process skills in Vertebrate Zoology. Previous studies have largely focused on conceptual achievement or general science process skills without developing psychometrically validated instruments tailored to specific biological content (Apriyani & Aini, 2025; Komisia et al., 2025). By addressing this gap, the present study demonstrates that contextualized assessment can provide more accurate information regarding prospective teachers' inquiry competencies while supporting authentic assessment practices in higher education. Nevertheless, further validation involving larger samples, multiple institutions, and additional Vertebrate Zoology topics is recommended to strengthen the generalizability and robustness of the instrument. Such continued refinement will contribute to the development of more reliable assessment systems capable of supporting inquiry-based biology education and preparing future teachers with the scientific competencies required for twenty-first-century learning.

4. CONCLUSION

This study successfully developed and validated an essay-based instrument for assessing science process skills (SPS) in the Vertebrate Zoology course, specifically within the Pisces superclass, for prospective biology teachers. The findings demonstrate that the instrument possesses satisfactory psychometric properties, including acceptable content validity, adequate construct validity, high item reliability, and a balanced distribution of item difficulty, indicating its potential as an authentic assessment tool for evaluating students' scientific inquiry competencies beyond conceptual knowledge. The application of the Rasch model also enabled the identification of two misfitting items requiring revision, highlighting the value of psychometric analysis in refining educational assessment instruments. Despite these promising results, the study is limited by the relatively small sample size, the involvement of participants from only two universities, and the instrument's focus on a single topic within Vertebrate Zoology, which may restrict the generalizability of the findings. Future research is therefore recommended to validate the instrument using larger and more diverse samples across multiple institutions, extend its application to other biology topics, and examine its relationship with broader educational outcomes, such as scientific literacy, critical thinking, and inquiry competence, in order to further strengthen its validity, reliability, and practical applicability in biology teacher education.

REFERENCES

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/https://doi.org/10.1177/0013164485451012

Anbiya, K., Muhibbudin, Khaldun, I., & Yusrizal. (2023). Integration of Problem-Based Learning Model with Guided Inquiry Worksheet to Enhance Scientific Process Skills and Critical Thinking Abilities. *Jurnal Penelitian Pendidikan IPA*, *9*(10), 8328–8334. https://doi.org/10.29303/jppipa.v9i10.4724

Apriyani, D., & Aini, F. Q. (2025). Design of Science Process Skills Instruments Reaction Rate Material: Rasch Model Approach. *Journal of Educational Sciences*, *9*(1), 157–172. https://doi.org/https://doi.org/10.31258/jes.9.1.p.157-172

Arifin, K., Nurijal, Jahidin, Munir, A., & Ede, S. G. (2025). The improvement of science process skills in prospective biology teacher students through the implementation of feedback A . Introduction. *BIO-INOVED: Jurnal Biologi-Inovasi Pendidikan*, *7*(1), 58–68. https://doi.org/10.20527/bino.v7i1.20049

Azzahra, M. Z., Nawahdani, A. M., & Falani, I. (2024). The relationship between science process skills and 21st century skills in science learning systematic literature review. *EduFisika: Jurnal Pendidikan Fisika*, *9*(3), 297–305. https://doi.org/10.59052/edufisika.v9i3.38451

Centinkaya, M., & Ozyurek, C. (2019). The Effect of Inquiry-Based Science Activities on Prospective Science Teachers' Scientific Process Skills. *International Online Journal of Education and Teaching*, *6*(1), 56–70.

Cinantya, G. C., & Sugiharto, B. (2026). Rasch model analysis: Grade 11 biology education questionnaire accuracy. *Biosfer: Jurnal Pendidikan Biologi*, *19*(1), 89–96. https://doi.org/10.21009/biosferjpb.58337

Cleaver, A., Crean, L., & Howitt, S. (2025). Student Perceptions of Cognitive , Psychomotor and Affective Learning in the Undergraduate Laboratory. *Biochemistry and Molecular Biology Education*, *53*(6), 610–618. https://doi.org/10.1002/bmb.70011

Farach, N., Kartimi, & Mulyani, A. (2021). Application of performance assessment in STEM-based biological learning to improve student's science process skills. *Journal of Physics: Conference Series*, *1806*(1). https://doi.org/10.1088/1742-6596/1806/1/012220

Fauziyah, S. L., Ismail, I., & Norra, B. I. (2020). Pengembangan Media Pembelajaran Scrapbook Untuk Meningkatkan Motivasi Belajar Pada Materi Vertebrata di Kelas X Pendidikan Biologi , Universitas Islam Negeri Walisongo Semarang PENDAHULUAN Kegiatan belajar mengajar merupakan kegiatan yang paling pokok dal. *Bioeduca : Journal of Biology Education*, *2*(1), 16–27.

Hafizi, N., Khalid, M., Yusof, I. J., Latif, A. A., & Jani, D. (2023). Psychometric properties of teacher classroom assessment literacy instrument using Rasch model analysis. *International Journal of Evaluation and Research in Education (IJERE)*, *12*(2), 638–646. https://doi.org/10.11591/ijere.v12i2.23607

Inayah, A. D., Ristanto, R. H., Sigit, D. V, & Miarsyah, M. (2020). Analysis of science process skills in senior high school students. *Universal Journal of Educational Research*, *8*(4 A), 15–22. https://doi.org/10.13189/ujer.2020.081803

Jannah, S. S. F., & Widyanti, E. (2024). Assessment of Knowledge Competency Achievement. *Journal of Scientific Research, Education and Technology*, *3*, 881–889. https://doi.org/10.58526/jsret.v3i2.429

Kadir, A., Zulqarnain, T., Takda, A., Assingkily, M. S., & Ahmad, M. (2025). Science Literacy Skills of Junior High School Students Based on the Nature of Science Literacy Test (NOSLiT) Instrument. *Jurnal Pendidikan IPA Indonesia*, *14*(1), 93–101. https://doi.org/10.15294/jpii.v14i1.15739

Komisia, F., Tukan, M. B., Yanni, D., Fallo, A., Jenisa, G., Hale, P., … Leba, U. (2025). Validity Analysis of a Science Process Skills Assessment Instrument for High School Students in Chemistry. *Jurnal Penelitian Pendidikan IPA*, *11*(12), 185–193. https://doi.org/10.29303/jppipa.v11i12.12860

Kostøl, K. B., Bøe, M. V., & Skår, A. R. (2023). Nature of Science in Norway's Recent Curricula Reform: Analysis of the Biology, Chemistry, and Physics Curricula. *Science and Education*, *32*(5), 1561–1581. https://doi.org/10.1007/s11191-022-00399-z

Kwok, S. (2018). Science Education in the 21 st Century. *Nature Astronomy*, *2*(7). https://doi.org/10.48550/arxiv.1806.08042

Moore, A. M., Fairhurst, P., Bennett, J. M., Harrison, C., Correia, C. F., Durk, J., … Harrison, C. (2024). Assessment and practical science : identifying generalizable characteristics of written assessments that reward and incentivise effective practices in practical science lessons. *International Journal of Science Education*, *46*(7), 643–669. https://doi.org/10.1080/09500693.2023.2253366

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*. Retrieved from https://timssandpirls.bc.edu/timss2019/international-results/

Nurhayati, H. O., Sartika, S. B., & Salim, A. (2025). Experimental Method Improves Psychomotor Skills in Natural Science Learning : Metode Eksperimental Meningkatkan Keterampilan Psikomotor dalam Pembelajaran Ilmu Pengetahuan Alam. *Indonesian Journal of Education Methods Development*, *20*(4), 1–8. https://doi.org/10.21070/ijemd.v20i4.900

Nurita, T., Fauziah, A. N. M., Sudibyo, E., & Mahdiannur, M. A. (2023). Exploring prospective teachers' science process skills: A voice from freshmen to senior of science education major. *AIP Conference Proceedings*, *2540*(1). https://doi.org/https://doi.org/10.1063/5.0105767

Purnamasari, L. (2016). *Hubungan Antara Makna Kerja dengan Keterlibatan Kerja pada Karyawan di*

Bandung. Universitas Pendidikan Indonesia.

Razali, A., Nor, M., & Mohd, S. (2020). Analysis of mastery of science process skills and 21 st century skills among secondary school students. *Jurnal Pendidikan Sains & Matematik Malaysia*, *10*(1), 16–20. https://doi.org/10.37134/JPSMM.VOL10.1.3.2020

Rozali, M. Z., Puteh, S., Amin, F., Yunus, N., & Hamdan, N. H. (2022). Reliability and Validity of Instrument on Academic Enhancement Support for Student-Athlete Using Rasch Measurement Model. *Asian Journal of University Education (AJUE)*, *18*(1). https://doi.org/10.24191/ajue.v18i1.17199

Sanches de Oliveira, G., van Es, T., & Hipólito, I. (2023). Scientific practice as ecological-enactive co-construction. *Synthese*, *202*(1), 1–33. https://doi.org/10.1007/s11229-023-04215-1

Sheptian, R., Arthur, R., & Wardhani, R. (2024). Psychometric analysis of mathematics readiness assessment tool for elementary school students: a rasch model approach. *Learning : Jurnal Inovasi Penelitian Pendidikan Dan Pembelajaran*, *4*(4), 1213–1220. https://doi.org/10.51878/learning.v4i4.3725

Sugiyono. (2019). *Metode Penelitian Pendidikan Pendekatan Kuantitatif, Kualitatif, dan R&D*. Bandung: Alfabeta.

Suliyanah, Adelia, B. D., Jauhariyah, M. N. R., Misbah, Mahtari, S., Saregar, A., & Deta, U. A. (2021). A bibliometric analysis of minimum competency assessment research with VOS viewer related to the impact in physics education on 2019-2020. *Journal of Physics: Conference Series*, *2110*(1). https://doi.org/10.1088/1742-6596/2110/1/012022

Thiagarajan. (1974). *Instructional development for training teachers of exceptional children*. Indiana: ERIC - Education Resources Information Center. Retrieved from https://eric.ed.gov/?id=ED090725

Ubaidillah, M., Marwoto, P., Wiyanto, Rusilowati, A., Subali, B., Mindyarto, B. N., & Isnaeni, W. (2022). Development of Habits of Mind Instruments in the Context of Basic Physics Practicum: EFA and Rasch Model. *Journal of Educational, Cultural and Psychological Studies (ECPS)*, *26*(December). https://doi.org/https://dx.doi.org/10.7358/ecps-2022-026-ubai

Ugli, U. K. K. (2024). Developing critical thinking, research, and problem-solving skills through historical contexts in improving natural science literacy of schoolchildren. *American Journal of Social Sciences and Humanity Research*, *4*(10), 361–366. https://doi.org/10.37547/ajsshr/volume04issue10-29

Windi, R. (2011). Uji Validitas dan Reliabilitas dalam Penelitian Epidemiologi Kedokteran Gigi. *J.K.G Unej*, *8*(1), 27–34.

Yudha Ikhwal Maryanto, Ganda Hijrah Selaras, Lufri, & Fitri Olvia Rahmi. (2026). Science Process Skills in Science Learning: A Literature Review. *Jurnal Penelitian Guru Indonesia*, *6*(2), 2158–2166. https://doi.org/https://doi.org/10.58578/tsaqofah.v6i2.8799

Yunita, L., & Mandasari, N. (2025). Pendidikan Sains Berorientasi Keterampilan Abad 21 dalam Konteks Pendidikan Tinggi: Review. *Panthera : Jurnal Ilmiah Pendidikan Sains Dan Terapan*, *5*(1), 40–49. https://doi.org/https://doi.org/10.36312/panthera.v5i1.334

Zahra, A., Nahadi, & Siswaningsih, W. (2025). Assessing Students ' Science Process Skills : Designing a Performance-Based Assessment Instrument for Saponification Lab Activities. *Jurnal Penelitian Pendidikan IPA*, *11*(7), 702–712. https://doi.org/10.29303/jppipa.v11i7.11757

Zhou, M. (2023). Significance of assessment in learning: The role of educational assessment tools. *Science Insights Education Frontiers*, *18*(2), 2881–2883. https://doi.org/10.15354/sief.23.co215

Zulfa, & Malik, A. (2025). The Development of 21st Century Skills through PSL Practicum and HOT Lab in Science Education. *Jurnal Pendidikan Fisika Dan Teknologi (JPFT)*, *11*(2). https://doi.org/https://dx.doi.org/10.29303/jpft.v11i2.8933

.