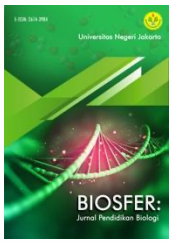
Biosfer: Jurnal Pendidikan Biologi

Journal homepage: http://journal.unj.ac.id/unj/index.php/biosfer



Polytomous items analysis with GPCM model on development of science process skills performance assessment in biology subject

Aulia Hermawati Ulfah*, Supahar

Educational Research and Evaluation, Graduate School, Yogyakarta State University, Indonesia

*Corresponding author: aulia.hermawati2021@student.uny.ac.id

ARTICLE INFO

Article history
Received: 28 January 2025
Revised: 27 March 2025
Accepted: 18 December 2025

Keywords:
Instrument Analysis
Item Response Theory
Performance Assessment
Science Process Skills

ABSTRACT

Assessment of science process skills can be conducted through tests or non-tests. However, observations, as one non-test technique, often face challenges related to space, time, and evaluator resources. A polytomous-type paper-and-pencil test was developed to address these issues. Item characteristic analysis is an essential aspect of instrument development; thus, this study reports the instrument analysis within the development of this assessment, using the item response theory approach. A total of 415 high school students in Sumedang completed 30 items depeloved. They came from four high schools selected based on the 2022 Asessmen Kompetensi Minimum (AKM) criteria. The test of 30 items had previously undergone successful content validity, construct validity, and reliability tests. Based on the item response theory assumptions, all items were eligible for analysis using this approach. Item fit testing revealed that the instrument suited the Generalized Partial Credit Model (GPCM). The analysis was performed using the R Program application. Results showed that all items (100%) were medium difficulty; 21 items (70%) had good discrimination and items (30%) exhibited high discrimination. 14 items were suggested for category reduction due to minimal delta differences or exceedingly low delta values. These findings emphasize item analysis to achieve high-quality instruments.

INTRODUCTION

Today, education requires implementing 21st-century learning to prepare a quality generation. In Indonesia, that is reflected in the implementation of the 2013 Curriculum (Putro et al., 2019) and continued with the Merdeka Curriculum. This is outlined in the Ministry of Education and Culture Regulation No. 5 of 2022, which sets graduate competency criteria covering attitudes, knowledge, and psychomotor domains. In biology learning, these three competencies are achieved through knowledge elements and science process skills elements. Experts have defined these process skills, which are commonly referred to as science process skills (Bass et al., 2009; D. J. Martin, 2009; Rezba et al., 1995). The number of science process skills (SPS) subcomponents varies across scholars. However, the most common basic skills are observing, classifying, communicating, inferring, predicting, and measuring, while interpreting and designing investigations were considered integrated skills. SPS can support students in developing their affective, cognitive, and psychomotor domains (Siahaan et al., 2020; Siswono, 2017). The development of SPS supports the success of biology learning, as scientific knowledge is fundamentally acquired through investigative processes (Arjaya & Suma, 2023; Chiappetta & Koballa, 2010; Ziraluo, 2021). In addition, SPS also contributes to students' success in higher education and the workforce (Irwanto, 2023; Mursali et al., 2024). This is because SPS reinforces essential 21st-century skills that are valuable for the future, such as problem-solving, critical thinking, creative tendencies, and higher-order thinking (Ekici & Erdem, 2020; Li et al., 2024). Therefore, teachers must integrate these skills into biology instruction.

SPS is essential for students in both developing countries, and its implementation has been in place since the 1980s (Mushani, 2021). SPS has received considerable attention in global science education due to its nature as transferable skills that allow students to emulate and demonstrate the behavior of scientists. However, research findings regarding students' SPS performance remain unsatisfactory (Baharom et al., 2020; Biswal & Behera, 2023; Kamarudin et al., 2022). Similar trends are observed in various regions of Indonesia. For instance, a study by Agus Kurniawan et al., (2020) showed that 55.2% of students' SPS were categorized as poor; research by Rifatul Mahmudah et al., (2019) indicated that 76% of students demonstrated low SPS; Harja & Sinaga (2021) reported that 75% of students had low SPS; and Sholihah et al. (2020) found that only 47% of the five SPS aspects assessed reached a satisfactory category. The low levels of SPS are attributed to teacher-centered instruction and the limited emphasis on SPS training (Agus Kurniawan et al., 2020; Baharom et al., 2020; Harja & Sinaga, 2021). Moreover, SPS assessment still tends to focus primarily on conceptual knowledge (Sholihah et al., 2020).

SPS can be enhanced through multiple strategies, including the integration of SPS into curriculum content, classroom instruction and assessment, student-centered learning approaches, and targeted SPS training (Gizaw & Sota, 2023). In Indonesia, Merdeka Curriculum has embedded SPS as a core process skill element within biology instruction, operationalized through student-centered pedagogical approaches. These approaches have demonstrated effectiveness in various international contexts, particularly through the implementation of methods such as Problem-Based Learning (PBL), STEM education, Inquiry-Based Instruction, and Guided Inquiry-Based Learning (Kasuga et al., 2022; Sari et al., 2020; ŞEN & VEKLİ, 2016). Each of these approaches includes specific instructional phases designed to foster SPS development. However, to maximize their impact, effective and appropriate assessment tools are also essential. Commonly adopted assessment techniques for SPS include performance-based assessments, utilizing both observation and written tests. These methods have been empirically shown to contribute positively to the development of students' SPS (Im et al., 2024; Srirahayu & Arty, 2019; Vo & Simmie, 2024).

This study assumes a similar role in developing a performing-based assessment instrument for SPS using written test format within the context of biology instruction. The instrument was specifically designed to address sommon challenges in assessing students' SPS during investigative activities, such as limited instructional time, large class sizes, and the extensive number of indicators that need to be evaluated (Zuhera & Habibah, 2017). The tests items were constructed to simulate authentic investigative scenarios, prompting students to respond based on their retained knowledge and skills following an inquiry experience. A similar assessment approach was previously developed by Prasetyo (2015) for evaluating inquiry-based learning activities in physics. The present study adapts and extends this approach to suit the context of biology learning. Initially, a semi-divergent format combining multiple-choice essay items was employed, drawing on the work of Subali (2009) and Supahar & Prasetyo (2015). However, based on the results of a small-scale pilot study, the assessment format was revised to a convergent format consisting solely of multiple-choice items. This adjustment was made in

response to students' tendency to leave the essay response blank. The instrument utilizes polytomous items, which have the advantage of enhancing student motivation, as each response step is awarded a score ranging from minimum to maximum (Gorgun & Bulut, 2021).

The development of the assessment instrument in this study is limited to the *Plantae* topic. Indonesia, as a country rich in biodiversity, holds great potential for integrating plant-based learning to enhance students' literacy on plant life, ultimately contributing to environmental conservation (Nurdini et al., 2020). Previous research has indicated that students' abilities to evaluate, classify, and compare medical plant species remain inadequate (Anwar et al., 2015). To support high-quality learning in the *Plantae* unit, this topic was selected as the focus for instrument development. Moreover, assessment instruments specifically designed for the *Plantae* topic remain relatively scarce.

Assessment instruments must be high-quality to accurately reflect students' abilities. An assessment is considered accurate if it minimizes errors (Azwar, 2022). Instrument quality is determined through validity, reliability, and item characteristic analysis (Allen & Yen, 1979; Azwar, 2022; Haryanto, 2020; Istiyono, 2020). Therefore, item analysis is crucial in instrument development. Such analysis can be conducted using classical test theory or modern test theory approaches (Istiyono, 2020; Mardapi, 2016). However, modern test theory, also known as item response theory (IRT), is preferred due to its ability to address classical test theory's weaknesses (Eaton et al., 2019; Sarea & Ruslan, 2019). Various item analysis models using IRT have been developed. 1PL, 2PL, and 3PL models can be used for dichotomous items. For polytomous items, models such as the Graded Response Model (GRM), Partial Credit Model (PCM), Generalized Partial Credit Model (GPCM), and Rating Scale Model (RSM) are available (de Ayala, 2009; Gruijter & Kamp, 2008). These models are chosen based on the scoring scale used. Polytomous scales are advantageous as they allow educators to analyze students' errors or shortcomings in selecting answers, unlike dichotomous scoring, which only considers correct or incorrect answers (A. Prasetya et al., 2019).

In recent years, item response theory (IRT) has gained attention in educational assessment research. This is evident from several studies, such as: Pongsophon & Jituafua (2021) analyzed an assessment for learning progression in botanical literacy using the Rasch Model; Ma et al., (2025) validated a five-tier diagnostic instrument for respiration and photosynthesis based on the Rasch model; Karoror and Jalmo (2022) applied the Rasch model to examine students' critical thinking skills on the topic of ecosystems; Andriani et al. (2021) used Rasch analysis to evaluate items in a four-tier multiple-choice test on the immune system; Prasetya & Pratama (2023) assessed item quality in a critical thinking skills test on the human digestive system using the Rasch model; and Paxinou et al., (2021) evaluating learning gains from a virtual reality laboratory experience using an IRT-based approach. Given these precedents, it is highly relevant to analyze SPS assessment items using the IRT approach. The present study reports items characteristic of the developed to assess students' SPS in biology education using IRT. Therefore, this study aims to evaluate the quality of the development assessment items. The results of the item analysis will serve as a basis for refining the instrument for broader implementation.

METHODS

Research Design

The assessment instrument in this study was developed using the ADDIE model (analyze, design, develop, implement, and evaluate) (Branch, 2009), combined with the instrument development steps by Mardapi (2016). This study focuses on a part, named items characteristic analysis, employing a quantitative descriptive research approach.

Population and Samples

The sample consisted of 415 grade XI students who had completed instruction on the *Plantae* topic in Biology. These students were selected using purposive sampling from four schools located in Sumedang Regency. School selection was based on the results of the 2022 Minimum Competency Assessment (*Assessmen Kompetensi Minimum*, AKM). The selected school represented three performance categories: one school categorized as below the minimum competency level, two schools at the minimum competency level, and one school above the minimum competency level. These criteria were applied to ensure a heterogeneous sample in terms of student ability, representative of the broader population of high school students in Sumedang Regency. The selected schools for this study were SMAS Al-Aqsha, SMAN 1 Sumedang, SMAN 1 Jatinangor, dan SMAS PGRI Parakan Muncang.

Instrument

The instrument analyzed in this study was a performance assessment designed to measure students' knowledge retention following an investigation or inquiry activity. The test employed convergent multiple-choice questions, utilizing multiple-response items. Students were instructed to select three out of five options that best reflected the knowledge they had gained from the investigation. The initial construct consisted of 36 items. Students could earn a maximum score of 4 by selecting all three correct options, with scores decreasing progressively to a minimum of 1 depending on the number of correct responses. The scoring criteria were established based on a redefined assessment rubric.

Test items were developed based on a test blueprint, which was constructed according to the SPS frameworks proposed by several experts. Variations exist in the formulation of the SPS indicator among these experts. Bryce (1990) proposed basic SPS indicators, including observing, recording data, manipulating, measuring, following procedures, and selecting procedures. In contrast, Rezba (1995) categorized basic SPS as observing, inferring, classifying, predicting, measuring, and communicating, while integrated SPS included identifying variables, operationally defining variables, formulating hypotheses, organizing data into a table, describing relationships between variables, constructing graphs, analyzing investigation results, collecting and processing data, designing investigations, and conducting experiments. On the other hand, Chiapetta and Koballa (2010) defined basic SPS as observation, classification, apace/time relations, using numbers, measuring, inferring, and predicting, whereas integrated SPS encompassed operational definitions, model formulation, variable control, data interpretation, hypothesizing, and experimenting.

Based on the frameworks proposed by various scholars, only eight indicators were selected for item development: observing, classifying, communicating, inferring, predicting, and measuring as basic SPS, as well as interpreting and designing investigations as integrated skills. Although these basic skills are expected to be well-developed among high school students (Khamhaengpol et al., 2024). Several studies have reported unsatisfactory performance in these areas (Harja & Sinaga, 2021; Rifatul Mahmudah et al., 2019; Sholihah et al., 2020; Yunita & Nurita, 2021). Moreover, inquiry-based instruction experienced by students during the *Plantae* unit predominantly engaged these eight indicators more than others. Consequently, the test items were constructed based on these selected indicators. Each indicator was further elaborated into sub-indicators, which were then used as the basis for item development.

Procedure

1. Analyze
   The first step in instrument development was analyzing needs through field data collection and literature review to determine instrument development requirements.
2. Design
   The product developed is in the form of a performance assessment instrument, so the design phase involves: 1) determining instrument specifications, including objectives and blueprint; 2) defining the scoring scale; and 3) establishing scoring criteria (Multiningsih, 2011). The goal was to produce a valid, reliable, and efficient instrument for measuring instrument science process skills in the biology subject. A blueprint was created based on curriculum analysis to identify the competencies students must achieve, followed by creating a matrix to guide item development and rubric formulation.
3. Develop
   The development phase is the phase of producing and measuring the eligibility of the instrument. This phase concluded with item writing and rubric development, content validation, limited trials, item analysis, and instrument refinement. The first draft consisted of 36 items, trialed with 20 students to assess the format's functionality. Content validation involved six experts, including assessment experts, Biology matter experts, and high school Biology teachers. Adjustments based on their feedback reduced the instrument to 35 items. After a broader trial with 415 students, construct validity was performed leading to a final set of 30 items analyzed using IRT.
4. Implement
   This phase is the implementation stage of using the instrument. The measurement results provided data for evaluating individual student abilities and identifying areas needing improvement.
5. Evaluate
   This phase assessed the instrument's quality before and after implementation. Researchers also identified successes and recommended improvements for future projects.

Data Analysis Techniques

Item and instrument analysis aims to evaluate how well the items and instruments provide information related to the measured variables (Samritin & Suryanto, 2016). Item Response Theory (IRT) analysis begins with assumption testing.

1. Assumption Testing in Item Response Theory (IRT)
   Assumption testing in IRT involves verifying unidimensionality, local independence, and invariance. The unidimensionality assumption can be tested using factor analysis by examining eigenvalues that form factors (Kartowaginaran et al., 2019). Local independence is satisfied if the instrument meets the unidimensionality assumption (Humbleton, 1991). Meanwhile, the invariance assumption is tested by estimating item parameters for two groups, including several parameters such as discrimination, difficulty, and guessing, which are presented in the scatterplot (Retnawati, 2016).
2. Item Fit Testing
   If the three assumptions of IRT are met, item fit testing against the IRT model follows. For this instrument, data were tested using the PCM or GPCM models, as the items used a partial credit scale. The model with the highest number of fit items was selected for further analysis. The criteria for a fit model specify that an item is considered fit if its infit and outfit values are between 0.75 and 1.3. Item falling outside this range are deemed misfit, as interpreted according to Zi Yan & Heene (2021), shown in the following Table 1.

Table 1.
Criteria for Misfit and their Interpretation

Infit or Outfit Value	Respons Pattern	Variation	Interpretation	Criteria
>1,3	Too haphazard	Too much	Unpredictable	Underfit
<0,75	Too determined	Too little	Guttman	Overfit

3. Items Characteristics Analysis
   PCM and GPCM are IRT models that analyze polytomous items with a partial credit scale. PCM is an extension of the Rasch model that involves a one-parameter logistic analysis (Ayala, 2009; Zi Yan & Heene, 2021). Besides that, GPCM extends PCM by incorporating two logistic parameters, providing more detailed information than PCM (van der Linden and Hambleton, 1997). The research data were analyzed, yielding fit results for the GPCM model. Therefore, item characteristics in this study were analyzed using GPCM. The study presents the results of testing two item characteristics: difficulty level and discrimination. The categories for the two logistic parameters are based on the criteria from Hambleton et al. (1991).

Table 2.
Criteria and Category of Logistic Parameter of GPCM

Logistic Parameter	Criteria	Category
Difficulty	b > 2	Hard
	-2 > b > 2	Medium difficulty
	< -2	Easy
Discrimination	a > 2	High
	0 ≤ a ≤ 2	Moderate
	a < 0	Low

RESULTS AND DISCUSSION

Before being analyzed using the IRT approach, the instrument in this study underwent content validation, construct validation, and reliability testing. Based on these tests, 30 items were deemed suitable for analysis using the IRT approach. Unlike classical test theory, IRT evaluates item quality by comparing the mean performance items against the evidence of group ability as estimated by the model (van der Linden & Hambleton, 1997). Through IRT analysis, the characteristics of the items can provide precise results at each level of the scoring scale, which is more effective than relying on a single reliability value for all test attributes (Reeve & Masse, 2004). According to Hambleton et al. (1991) three

assumptions must be fulfilled before conducting IRT analysis: unidimensionality, local independence, and invariance.

Assumption Testing in Item Response Theory

Unidimensionality

The unidimensionality assumption can be tested if the data meet the prerequisites for factor analysis, specifically when the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy (MSA) value is at least 0.5. Based on the estimation result, the KMO MSA value for the respondents' answers was 0.924, indicating that factor analysis could proceed. These results are shown in Table 3.

Table 3.
Kaiser-Meyer-Olkin tes

	MSA
Overall MSA	0.924

A unidimensional test through factor analysis is done by looking at the eigenvalue on the screen plot. The scree plot for this research data is shown in Figure 1.

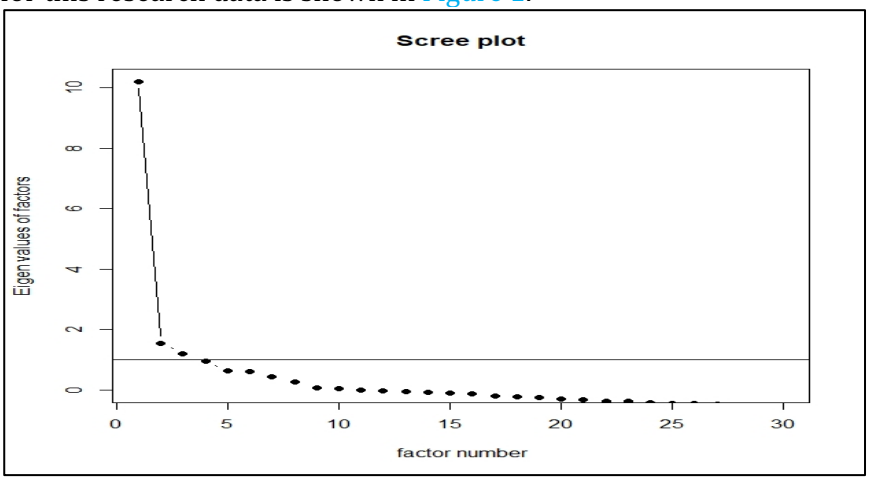


Figure 1. Scree plot of Science Process Skills Performance Assessment Instrument

Figure 1 illustrates the scree plot for the data, demonstrating unidimensionality as there is a single dominant factor compared to the others. This is evident from one point at the peak that is significantly higher than the other points below it. Additionally, there are three points above the line (eigenvalue >1). The first factor has the highest eigenvalue (more than 10), while the second and subsequent factors have eigenvalues < 2.

Local Independence

Local independence indicates that the data are independent when participants' responses to test items remain consistent. A participant's response to one item is not influenced by their response to other items or those of other participants. The assumption of local independence is satisfied if the unidimensionality assumption is met (DeMars, 2010; Hambleton et al., 1991). Since the data in this study are unidimensionality, the assumption of local independence is fulfilled.

Invariance

Two invariances must be met: item parameter invariance and ability parameter invariance. The results of the invariance test are presented in Figures 2 and 3.

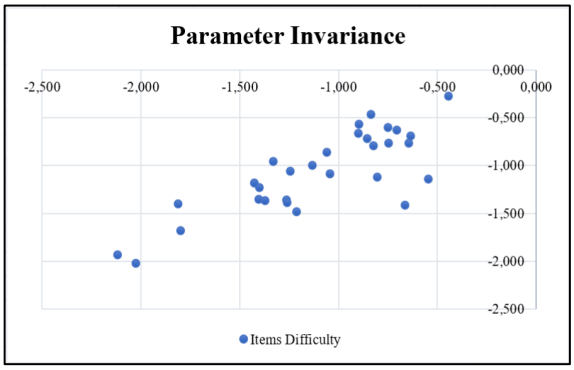


Figure 2. Parameter Invariance of Odd and Even Participants in Science Process Skills Performance Assessment Instrument

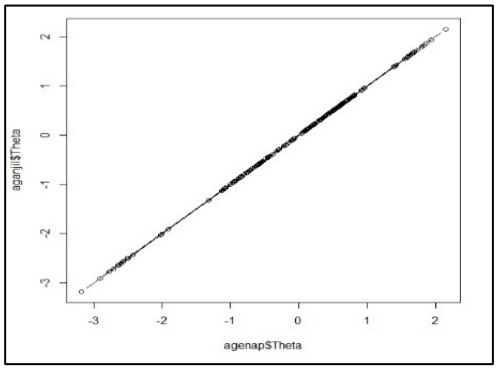


Figure 3. Invariance of Odd and Even Participants' Abilities in Science Process Skills Performance Assessment Instrument

A scatterplot indicates invariant data if the points on the plot align closely with a line passing through the origin with a slope of 1 (Retnawati, 2014). In Figures 2 and 3, the data points tend to be closely around the line with a slope 1, indicating no significant variation in item and ability parameters. Therefore, the assumption of invariance is satisfied.

Item Fit Testing

Item fit testing evaluates how well an item functions in measurement and its contribution to the overall instrument construction (Zi Yan & Heene, 2021). This is observed through the participants' response patterns to test items. This study examined student responses using the Partial Credit Model (PCM) and the Generalized Partial Credit Model (GPCM), as the scoring system employed a partial credit scale. Item fit estimation can be assessed through the Chi-square value, but this study opted to use infit and outfit values due to the Chi-square's sensitivity to large datasets. Based on the criteria, an item is considered fit if its infit or outfit values fall within the range of 0.75 to 1.3. An infit or outfit value >1.3 indicates underfit (variation exceeds the model's expectation), while a value < 0.75 indicates overfit (variation is below the model's expectation) (Zi Yan dan Heene, 2021). The item fit estimation was conducted using R software, and the results for each item are shown in Table 4.

Table 4.
Fit Items of Science Process Skills Performance Assessment Instrument with PCM and GPCM Models

Items Codes	Items Number	PCM			GPCM		
		outfit	infit	Exp.	outfit	infit	Exp.
Item_1	1	1.153	1.173	Fit	0,995	0,993	Fit
Item_2	2	1.100	1.113	Fit	1	0,995	Fit
Item_4	3	0.970	1.065	Fit	0,943	1,011	Fit
Item_5	4	0.977	0.986	Fit	1,003	1,002	Fit
Item_6	5	1.267	1.251	Fit	0,991	0,981	Fit
Item_7	6	1.310	1.249	Mix	0,958	0,973	Fit
Item_8	7	1.070	1.085	Fit	0,987	0,989	Fit
Item_9	8	1.195	1.196	Fit	0,991	0,988	Fit
Item_10	9	0.996	0.996	Fit	0,979	0,981	Fit
Item_11	10	0.899	0.929	Fit	0,967	0,986	Fit
Item_12	11	0.854	0.903	Fit	0,924	0,976	Fit
Item_13	12	0.844	0.854	Fit	0,966	0,968	Fit
Item_14	13	0.919	0.973	Fit	0,941	0,991	Fit
Item_15	14	0.987	1.017	Fit	0,965	0,991	Fit
Item_16	15	1.300	1.252	Fit	0,999	0,982	Fit
Item_17	16	1.119	1.116	Fit	0,978	0,977	Fit
Item_18	17	1.060	1.055	Fit	1,015	0,987	Fit
Item_19	18	1.116	1.105	Fit	1,011	0,991	Fit
Item_20	19	1.149	1.157	Fit	0,99	0,991	Fit
Item_21	20	1.185	1.164	Fit	1,01	0,986	Fit
Item_22	21	0.765	0.765	Fit	0,951	0,954	Fit
Item_23	22	0.701	0.705	Overfit	0,934	0,944	Fit

Items Codes	Items Number	PCM			GPCM		
		outfit	infit	Exp.	outfit	infit	Exp.
Item_24	23	0.707	0.710	Overfit	1,053	0,865	Fit
Item_25	24	0.724	0.721	Overfit	0,973	0,935	Fit
Item_27	25	0.747	0.747	Overfit	0,949	0,946	Fit
Item_28	26	0.753	0.723	Mix	1,452	0,803	Fit
Item_29	27	0.750	0.754	Fit	0,973	0,9	Fit
Item_31	28	0.733	0.741	Overfit	0,914	0,905	Fit
Item_32	29	0.693	0.695	Overfit	0,914	0,923	Fit
Item_33	30	0.727	0.737	Overfit	1,105	0,855	Fit

Based on the data in Table 3, the item fit estimation for the PCM model revealed seven overfit items and one mixed item. In contrast, the GPCM model showed that all items were fit. Overfit items indicate that response variation for those items was below the model's expectation. In contrast, a mixed item suggests that one of the infit or outfit values did not fit the model. Non-fit items may result from various factors, such as students' carelessness in answering questions (Ardiyanti, 2016), defects in the items themselves (e.g. low discrimination values), or items measuring unrelated abilities, leading to inconsistent or atypical response patterns (Faradillah & Febriani, 2021; Reise, 1982). Consequently, further IRT analysis was conducted using the GPCM model.

Item Characteristics

GPCM analysis yielded two logistic parameters: item difficulty and item discrimination. The results of these analyses are presented in Table 5. The location value indicates the difficulty level of an item, while the discrimination level is represented by a-value.

Table 5.
Logistic Parameter Values for the Science Process Skills Performance Assessment Instrument

Item number	a	b1	b2	b3	Location	Item Number	a	b1	b2	b3	Location
Item_1	0,597	-4,096	-2,064	0,385	-1,966	Item_17	0,649	-2,581	-1,900	2,398	-1,122
Item_2	0,718	-2,647	-1,801	2,133	-1,160	Item_18	0,716	-2,171	-1,816	2,288	-1,002
Item_4	0,733	-2,074	-1,224	-0,166	-1,167	Item_19	0,601	-3,538	-1,696	2,659	-1,205
Item_5	0,905	-2,217	-1,134	1,905	-0,752	Item_20	0,524	-4,328	-0,894	2,708	-0,861
Item_6	0,506	-1,519	-3,246	1,672	-1,478	Item_21	0,510	-2,773	-1,757	3,009	-0,913
Item_7	0,461	-3,079	-1,526	-1,027	-1,824	Item_22	1,864	-2,350	-1,278	0,821	-1,164
Item_8	0,688	-2,477	-1,325	1,635	-0,911	Item_23	2,128	-2,518	-1,209	0,838	-1,157
Item_9	0,558	-2,371	-1,956	1,446	-1,203	Item_24	2,243	-2,724	-1,020	0,152	-1,068
Item_10	0,851	-1,963	-1,154	1,134	-0,805	Item_25	2,303	-2,666	-1,435	0,779	-1,382
Item_11	0,953	-1,611	-1,781	1,523	-1,000	Item_27	2,064	-2,309	-1,076	0,887	-1,013
Item_12	0,976	-2,543	-0,603	0,049	-0,911	Item_28	2,196	-3,189	-0,844	-0,149	-1,005
Item_13	1,149	-1,916	-1,423	1,420	-0,969	Item_29	2,006	-2,302	-0,857	0,574	-0,859
Item_14	0,871	-1,674	-1,503	0,371	-1,056	Item_31	2,324	-2,675	-0,651	0,728	-0,684
Item_15	0,793	-2,099	-1,612	0,328	-1,237	Item_32	2,639	-2,493	-0,973	0,801	-0,964
Item_16	0,502	-2,364	-2,463	0,833	-1,557	Item_33	2,205	-2,979	-0,922	0,094	-1,011

All items in this analysis fall within the medium difficulty category, as their location values range from -1.966 to -0.684. According to Hambleton (1991: 13), an item is classified as easy if its location value is less than -2, medium difficulty between -2 and 2, and difficult if it exceeds 2. Regarding

discrimination, most items in this instrument exhibit good discrimination, with a-values in the range of 0-2, encompassing 21 items. The remaining items display high discrimination, with a value exceeding 2 (Hambleton et al., 1991). Extremely high or low discrimination values are considered suboptimal, as such items may fail to distinguish students' abilities effectively (Zainul & Nasoetion, 1997). Consequently, items with high discrimination values should be reviewed, revised, and re-evaluated until they achieve an acceptable discrimination range.

Table 6.
Frequency of Items Based on Difficulty Level and Discrimination According to Hambleton et al. (1991)

Difficulty				Discrimination			
Criteria	**Location Values**	**Frequency**	**Items Code**	**Criteria**	**a-Values**	**Frequency**	**Items Code**
Hard	b > 2	-	-	High	a > 2	9	Item_23 to Item_ 33
Medium difficulty	-2 > b > 2	30	(All items)	Moderate	0 ≤ a ≤ 2	21	Item_1 to Item_22
Easy	< -2	-	-	Low	a < 0	-	-

The logistic parameter values in Table 5. form the basis for projecting item characteristics onto the Item Characteristic Curve (ICC). The ICC illustrates the probability of students answering correctly based on their ability levels for each scoring category. The ICC for the items in this study is shown in Figure 4.

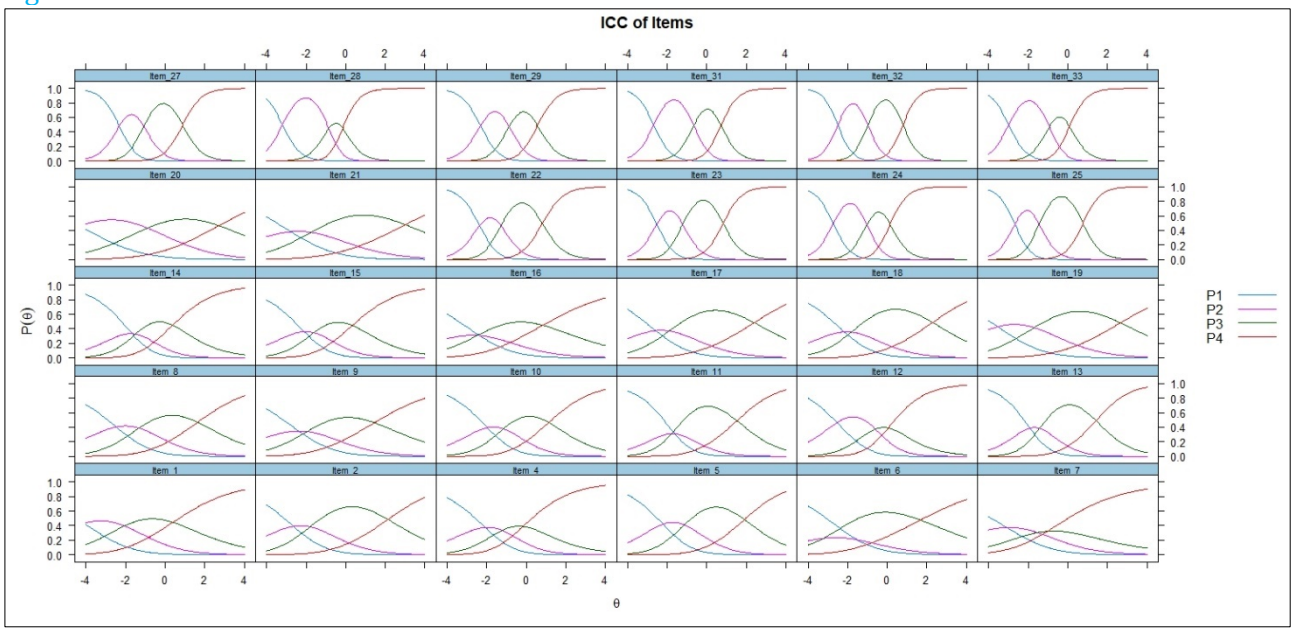


Figure 4. Item Characteristic Curve (ICC) of the Science Process Skills Performance Assessment Instrument

Each curve in Figure 4. features four colored lines, each representing a scoring category from 1 to 4. P symbolizes the probability of a student achieving a correct response. Specifically: P1 (blue line) represents the probability of scoring 1; P2 (purple line) represents the probability of scoring 2; P3 (green line) represents the probability of scoring 3. P4 (red line) represents the probability of scoring 4. The scale from -4 to 4 on the horizontal axis represents the logit scale, which is assumed to reflect students' ability levels, ranging from low to high. The intersections of the lines indicate the delta (b) values between categories, shown in Table 5. Each item, with its four scoring categories, has three delta (b) values.

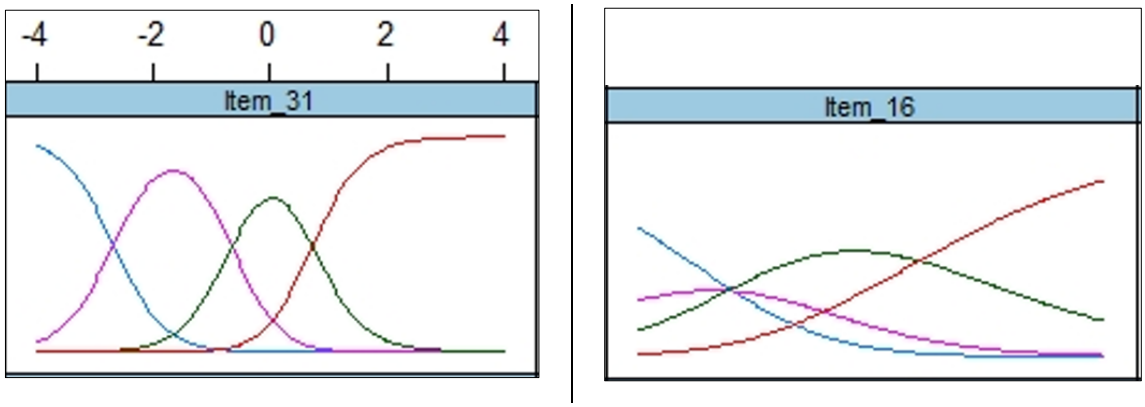


Figure 5. Item Characteristic Curve (ICC) Of Item_31 and Item_16

Thus, the ICC for item_31 indicates that category 1 is most likely achieved by students with ability levels below -2.675, students with ability levels likely achieve category 2 between -2.675 and -0.651., category 3 is most likely achieved by students with ability levels between -0,651 and 0,728., Students with ability levels above 0.728 most likely achieve category 4. Considering the delta (b) values, this item can be categorized as easy because students only need an ability level of -2.675 to achieve a score of 1, and a maximum score can be achieved with an ability level of just 0,728. The classification of student abilities based on the logit scale is presented in Table 7, following the criteria outlined by Hikamudin (2017).

Table 7.
Student Ability Criteria Based on Logit Scale

Range	Category
$\theta \leq -3$	Very Low
$-3 < \theta < -1$	Low
-1 < θ 1	Moderate
$1 < \theta < 3$	High
$\Theta > 3$	Very High

The graph for item_31 represents an ideal item characteristic curve, as its delta values are sequentially ordered. lower categories have a high probability of being selected by students with lower abilities, while higher categories are more likely to be selected by students with higher abilities. Other items with similar characteristics include Item_2, Item_4, Item_5, Item_8, Item_10, Item_12, Item_19, Item_22, Item_23, Item_24, Item_25, Item_27, Item_28, Item_29, Item_32, dan Item_33.

On the other hand, some items exhibit two reversed transition points due to unordered delta values. For instance, in item_16, as shown in Table 5, the first delta (b1) is -2,364, while the second delta (b2) is -2.463., since b1 is greater than b2, the transition points are reversed (de Ayala, 2009). This implies that achieving a score of 2 is easier than achieving a score of 1 for this item. Other items with similar reversed characteristics include Item_6, Item_11, and Item_16.

Furthermore, the b1 and b2 values of Item_16 are very close, resulting in overlapping transition points. Since the intersections are nearly identical, the scoring categories can be reduced to three (Dogan, 2018). Category 1 corresponds to the blue line, Category 2 to the green line, and Category 3 to the red line. This is further supported by the purple line's position, which overlaps beneath the intersection of the blue and green lines. Based on this observation, it is suggested to reduce the number of categories for other items with similar characteristics, including Item_6, Item_7, Item_9, Item_10, Item_11, Item_12, Item_13, Item_14, Item_15, Item_17, dan Item_18.

Additionally, for items with delta values less than -4, it is suggested to reduce the number of categories to three. Such items are too easy, making scoring categories more appropriate. Items with this characteristic include Item_1 and Item_20. Based on the characteristics of each item, several items were revised to meet the requirements. A total of 14 items now use a scale ranging from 1 to 3, while 16 items use a scale ranging from 1 to 4. The revised scoring categories or scales are presented in Table 8.

Table 8.
Revised Scored Scales for the Science Process Skills Performance Assessment Instrument

Item Code	Scale	Item Code	Scale
Item_1	1, 2, 3	Item_17	1, 2, 3
Item_2	1, 2, 3, 4	Item_18	1, 2, 3
Item_4	1, 2, 3, 4	Item_19	1, 2, 3, 4
Item_5	1, 2, 3, 4	Item_20	1, 2, 3
Item_6	1, 2, 3	Item_21	1, 2, 3, 4
Item_7	1, 2, 3	Item_22	1, 2, 3, 4
Item_8	1, 2, 3, 4	Item_23	1, 2, 3, 4
Item_9	1, 2, 3	Item_24	1, 2, 3, 4
Item_10	1, 2, 3	Item_25	1, 2, 3, 4
Item_11	1, 2, 3	Item_27	1, 2, 3, 4
Item_12	1, 2, 3	Item_28	1, 2, 3, 4
Item_13	1, 2, 3	Item_29	1, 2, 3, 4
Item_14	1, 2, 3	Item_31	1, 2, 3, 4
Item_15	1, 2, 3	Item_32	1, 2, 3, 4
Item_16	1, 2, 3	Item_33	1, 2, 3, 4

Instruction and assessment are inherently interconnected components of the educational process (Bichi et al., 2019; Kriswantoro et al., 2021). Teachers must employ a variety of assessment methods to obtain comprehensive information about students' competencies (Williams-McBean, 2025). Assessing SPS annable teachers to evaluate the extent to which students have understood and mastered the procedures they practice. According to Ilma et al., (2020) SPS are positively correlated with student learning outcomes. The assessment instrument developed in this study was based on a scientific approach, which aims to promote the enhancement of students' SPS. Previous research has demonstrated that performance assessment designed around inquiry-based learning, STEM education, or experimental activities is effective in improving students' SPS (Farach et al., 2021; Minalisa et al., 2019; Srirahayu & Arty, 2019). By engaging students in scientific activities and fostering the use of SPS, their understanding of Biologycal concepts can become more profound (Güler & Şahin, 2019). Furthermore, the application of SPS ensures that science instruction aligns with the fundamental goals of science education itself (Sarah et al., 2020).

In addition to considering the instructional approach in instrument development, the format of the test items must also be carefully selected to ensure optimal assessment of students' abilities (Slepkov et al., 2021). Polytomous items are generally recommended because they can provide more detailed information (Tu, 2017). In Biology assessments, for example, polytomous items are often used in inquiry-based evaluations or practical assessments, where items are scored using a partial credit scale. This means that any student response, even if only partially correct, is awarded a score, which emphasizes the importance of clearly communicating the scoring system to students before test administration. When students are informed that the items follow a polytomous format, they may be more motivated to maximize their efforts, knowing that their performance can yield a range of scores rather than a binary outcome (Gorgun & Bulut, 2021). In contrast, dichotomous items may lead students to guess answers without fully engaging their knowledge or skills. Therefore, the development of multiple-choice tests should aim to allow students to experience a reasonable degree of success, thereby supporting motivation and accurate assessment (Butler, 2018).

A well-constructed test format must be accompanied by optimal item analysis. IRT is recommended for analysing polytomous items. The IRT approach allows for more accurate estimation of both item parameters and student abilities compared to classical test theory (Chen et al, 2021). Moreover, IRT is superior in minimizing score misinterpretation, reducing bias, and supporting the implementation of computerized adaptive testing (CAT) (Soland, 2024; Thomas, 2019). Polytomous items consist of tiered score categories, each of which requires detailed characterization. The Generalized Partial Credit Model (GPCM) is suitable for estimating polytomous items, as it accommodates ordinal scores with step difficulties that do not necessarily follow a sequential order (Retnawati et al., 2017). In GPCM analysis, a particular step may be more difficult than the next, or vice versa. Through polytomous scoring, an evaluator can identify specific errors or deficiencies in the student's reasoning processes—an advantage not offered by dichotomous scoring (A. Prasetya et al., 2019).

In addition to the importance of selecting an appropriate scoring system, the composition of items within a test set also requires careful consideration. Ideally, an assessment instrument should comprise easy, moderate, and difficult items, with the moderate-level items positioned at the peak of the normal distribution (Sahin & Yildirim, 2018). According to Kunandar (2013), the recommended proportions are 25% easy items, 50% medium difficulty items, and 25% difficult items. While the current instrument cannot be considered ideal, it remains usable, as it predominantly includes medium difficulty items that can extract information from test-takers across low, medium, and high ability levels. This differs from instruments comprising only easy items, which cannot capture information from high-ability students, or only difficulty items, which fail to capture information from items from low-ability students. However, the determination of item proportions should align with the measurement objectives. For instance, if the aim is to select a few individuals from a large pool, the test should include predominantly difficult items. Conversely, if there is a shortage of test-takers, the test should feature predominantly easy items.

CONCLUSION

Based on the analysis using the GPCM model, all items were categorized as having medium difficulty, with difficulty level falling within the range of $-2 < b < 2$. Regarding discrimination, 21 items demonstrated good discrimination, while 9 items exhibited high discrimination. Items with high discrimination require further review and revision. A total of 14 items were proposed to be revised to a 1-2-3 scoring scale. Among these, 12 items exhibited very small delta differences, and 2 items had their first delta values below -4. Consequently, only 7 items were deemed ready for immediate use in measurement, while 23 items required revision. This indicates that developing a high-quality instrument is a complex process that demands meticulous design and thorough testing. Therefore, it is recommended that instrument development consider all aspects, including content, item types and formats, scoring scales, rubrics, language accuracy, and analytical techniques. Involving experts from diverse fields is crucial for improving instrument quality. Finally, this study involved 415 students, which should be expanded to produce analyses on a larger scale.

Beyond its practical implications, this study also offers significant theoretical and scientific contributions. Theoretically, it contributes to the body of knowledge on assessment analysis through the application of IRT. The use of IRT enhances the understanding of item characteristics, particularly for polytomous items, and demonstrates how theoretical frameworks can be effectively applied in educational settings. Scientifically, the study provides an alternative approach to measuring SPS. The use of a quantitative approach supported by R programming adds value by increasing the accuracy and reliability of science assessment instruments. Ultimately, the development of this instrument supports evidence-based decision-making in science education.

Limitations and Future Works

Although the GPCM analysis has provided insights into item difficulty and discrimination levels, several limitations remain in the conclusions drawn. First, the recommended revisions to the scoring scale for 14 items have not yet been further tested to evaluate their impact on the score reliability and interpretation. Second, only seven items were deemed immediately usable, indicating that the instrument still requires further refinement. Third, the instrument has not yet been developed in a digital format.

This instrument may serve as a complementary or comparative tool for other types of SPS assessment in Biology education. The researchers recommend involving a broader and more diverse population in future studies—both geographically and the terms of students' proficiency levels—to examine the stability of item characteristics. Given the researchers' current limitations in developing a digital-based assessment, future research could focus on the digital adaptation of this instrument to improve the efficiency of analysis. Moreover, future studies may also incorporate qualitative approaches to explore the underlying reasons behind students' responses.

REFERENCES

Agus Kurniawan, D., Putri Wirman, R., Wulan Dari, R., & Yuhanis, E. (2020). Description of student science process skills on temperature and heat practicum. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *24*(1), 88–101. https://doi.org/10.21831/pep

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Cole Publishing Company.

Andriani, F., Indrowati, M., & Sugiharto, B. (2021). Analysis items of the four-tier immune system multiple choice test instrument using Rasch model. *Biosfer*, *14*(1), 99–119.

https://doi.org/10.21009/biosferjpb.18020
Anwar, F., Kanwal, S., Shabir, G., Alkharfy, K. M., & Gilani, A. H. (2015). Antioxidant and antimicrobial attributes of different solvent extracts from leaves of four species of mulberry. International journal of pharmacology, 11(7), 757-765. https://doi.org/10.3923/ijp.2015.757.765
Ardiyanti, D. (2016). Aplikasi model rasch pada pengembangan skala efikasi diri dalam pengambilan keputusan karier siswa. *Jurnal Psikologi*, *43*(3), 248–263. https://doi.org/10.22146/jpsi.17801
Arjaya, I. B. A., & Suma, K. (2023). Problems of biology learning and evaluation analysis at the cipp model-based higher education level. *Biosfer*, *16*(1), 152–167. https://doi.org/10.21009/biosferjpb.26835
Azwar, S. (2022). *Reliabilitas dan validitas*. Pustaka Belajar.
Baharom, M. M., Atan, N. A., Rosli, M. S., Yusof, S., & Hamid, M. Z. A. (2020). Integration of science learning apps based on Inquiry-Based Science Education (IBSE) in enhancing students' Science Process Skills (SPS). *International Journal of Interactive Mobile Technologies*, *14*(9), 95–109. https://doi.org/10.3991/ijim.v14i09.11706
Bass, J. E., Contant, T. L., & Carin, A. A. (2009). *Teaching science as inquiry* (8th ed.). Allyn & Bacon.
Bichi, A. A., Ibrahim, R. H., & Ibrahim, F. B. (2019). Assessment of students' performances in biology: Implications for measurements and evaluation of learning. *Journal of Education and Learning (EduLearn)*, *13*(3), 301–308. https://doi.org/10.11591/edulearn.v13i3.12200
Biswal, S., & Behera, B. (2023). Enhancing science process skills through inquiry-based learning: a comprehensive literature review and analysis. *International Journal of Science and Research (IJSR)*, *12*(8), 1583–1589. https://doi.org/10.21275/sr23817121415
Branch, R. M. (2009). *Instructional Design: The ADDIE Approach*. Springer Science & Business Media.
Bryce, J., Toole, M. J., Waldman, R. J., & Voigt, A. N. N. (1992). Assessing the quality of facility-based child survival services. *Health policy and planning, 7*(2), 155-163.
Butler, A. C. (2018). Multiple-choice testing in education: are the best practices for assessment also good for learning? In *Journal of Applied Research in Memory and Cognition* (Vol. 7, Issue 3, pp. 323–331). Elsevier Inc. https://doi.org/10.1016/j.jarmac.2018.07.002
Chiappetta, E. L., & Koballa, T. R. (2010). *Science instruction in the middle and secondary school* (7th ed.). Pearson Education.
Chen, Z., Li, J., Wang, B., & Xu, B. (2025). Probabilistic normalization conditions of polytomous knowledge structures. *Communications in Statistics-Theory and Methods, 54*(15), 4877-4895. https://doi.org/10.1080/03610926.2024.2430735
de Ayala, R. J. (2009). *The theory and practice of item response theory*. The Guilford Press.
DeMars, C. (2010). *Item response theory*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001
Dogan, E. (2018). An application of the partial credit IRT model in identifying benchmarks for polytomous rating scale Instruments. *Practical Assessment, Research & Evaluation*, *23*(7), 1–10. https://doi.org/10.7275/1cf3-aq56
Eaton, P., Johnson, K., Frank, B., & Willoughby, S. (2019). Classical test theory and item response theory comparison of the brief electricity and magnetism assessment and the conceptual survey of electricity and magnetism. *Physical Review Physics Education Research*, *15*(1). https://doi.org/10.1103/PhysRevPhysEducRes.15.010102
Ekici, M., & Erdem, M. (2020). Developing science process skills through mobile scientific inquiry. *Thinking Skills and Creativity*, *36*. https://doi.org/10.1016/j.tsc.2020.100658
Farach, N., Kartimi, & Mulyani, A. (2021). Application of performance assessment in STEM-based *Biologi*cal 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
Faradillah, A., & Febriani, L. (2021). Mathematical trauma students' junior high school based on grade and gender. *Infinity Journal*, *10*(1), 53–68. https://doi.org/10.22460/infinity.v10i1.p53-68
Gizaw, G. G., & Sota, S. S. (2023). Improving science process skills of students: a review of literature. *Science Education International*, *34*(3), 216–224. https://doi.org/10.33828/sei.v34.i3.5
Gorgun, G., & Bulut, O. (2021). A polytomous scoring approach to handle not-reached items in low-stakes assessments. *Educational and Psychological Measurement*, *81*(5), 847–871. https://doi.org/10.1177/0013164421991211
Gruijter, D. N. de, & Kamp, L. J. T. van der. (2008). *Statistical test theory for the behavioral sciences*. Taylor & Francis Group.
Güler, B., & Şahin, M. (2019). Using inquiry-based experiments to improve pre-service science teachers'

science process skills. *International Journal of Progressive Education*, *15*(5), 1–18. https://doi.org/10.29329/ijpe.2019.212.1
Hambleton, R. K., Swaminatan, H., & Rogers, J. H. (1991). *Fundamental of item response theory*. Sage Publications, Inc.
Harja, M., & Sinaga, P. (2021). Evaluation of science process skills of high school students in Tapaktuan City on static fluid material. *Journal of Physics: Conference Series*, *1806*(1). https://doi.org/10.1088/1742-6596/1806/1/012016
Haryanto. (2020). *Evaluasi pembelajaran (konsep dan manajemen)*. UNY Press.
Hikamudin, E. (2017). Estimasi kemampuan siswa dalam ujian nasional menggunakan metode bayes. *Penelitian Kebijakan Pendidikan*, *10*(3). https://doi.org/10.24832/jpkp.v10i2.171
Ilma, S., Al-Muhdhar, M. H. I., Rohman, F., & Saptasari, M. (2020). The correlation between science process skills and biology cognitive learning outcome of senior high school students. *JPBI (Jurnal Pendidikan Biologi Indonesia)*, *6*(1), 55–64. https://doi.org/10.22219/jpbi.v6i1.10794
Im, R., Iwayama, T., & Osa, M. (2024). Assessing the science process skills of chemistry high school teachers: A case study in Cambodia. *Universal Journal of Educational Research*, *3*(3), 235–244. https://doi.org/10.17613/z0c2-8v50
Irwanto, I. (2023). Improving preservice chemistry teachers' critical thinking and science process skills using research-oriented collaborative inquiry learning. *Journal of Technology and Science Education*, *13*(1), 23–35. https://doi.org/10.3926/jotse.1796
Istiyono, E. (2020). *Pengembangan instrumen penilaian dan analisis hasil belajar Fisika dengan teori tes klasik dan modern* (2nd ed.). UNY Press.
Kamarudin, N., Wahida, M., & Ahrari, S. (2022). Exploring basic and integrated science process skills and their impact on science achievement among university students. *Journal of Public Administration and Governance*, *12*(4S), 74. https://doi.org/10.5296/jpag.v12i4s.20572
Karoror, I., & Jalmo, T. (2022). Profile of critical thinking ability in ecosystem materials using the Rasch model. *Jurnal Penelitian Pendidikan IPA*, *8*(3), 1599–1604. https://doi.org/10.29303/jppipa.v8i3.1394
Kasuga, W., Maro, W., & Pangani, I. (2022). Effect of problem-based learning on developing science process skills and learning achievement on the topic of safety in our environment. *Journal of Turkish Science Education*, *19*(3), 872–886. https://doi.org/10.36681/tused.2022.154
Khamhaengpol, A., Nokaew, T., & Chuamchaitrakool, P. (2024). Development of STEAM activity "Eco-Friendly Straw" based science learning kit to examine students' basic science process skills. *Thinking Skills and Creativity*, *53*. https://doi.org/10.1016/j.tsc.2024.101618
Kriswantoro, Kartowagiran, B., & Rohaeti, E. (2021). A critical thinking assessment model integrated with science process skills on chemistry for senior high school. *European Journal of Educational Research*, *10*(1), 285–298. https://doi.org/10.12973/EU-JER.10.1.285
Kunandar. (2013). *Penilaian autentik: Penilaian hasil belajar peserta didik Kurikulum 2013*. RajaGrafindo Persada.
Li, X., Zhang, Y., Yu, F., Zhang, X., Zhao, X., & Pi, Z. (2024). Do science teachers' believes related to inquiry-based teaching affect students' science process skills? Evidence from a multilevel model analysis. *Disciplinary and Interdisciplinary Science Education Research*, *6*(1). https://doi.org/10.1186/s43031-023-00089-y
Ma, H., Liu, W., & Li, G. (2025). Development and application of a five-tier diagnostic test to assess misconceptions on respiration and photosynthesis among senior high school students in Mainland China. *Research in Science Education*. https://doi.org/10.1007/s11165-025-10232-6
Mardapi, D. (2016). *Pengukuran, penilaian, dan evaluasi pendidikan* (2nd ed.). Parama publishing.
Martin, D. J. (2009). *Elementary science methods a constructivist approach* (5th ed.). Wadsworth Cengage Learning.
Minalisa, M., Festiyed, & Ratnawulan. (2019). The development of performance assessment of inquiry-based learning (IBL) to improve student's science process skill of class XI Senior High School 1 Bayang. *Journal of Physics: Conference Series*, *1185*(1). https://doi.org/10.1088/1742-6596/1185/1/012134
Mulyatiningsih, E. (2011). *Riset Terapan Bidang Pendidikan dan Teknik*. UNY Press.
Mursali, S., Sri Hastuti, U., Zubaidah, S., & Rohman, F. (2024). Guided inquiry with Moodle to improve students' science process skills and conceptual understanding. *International Journal of Evaluation and Research in Education*, *13*(3), 1875–1884. https://doi.org/10.11591/ijere.v13i3.27617
Mushani, M. (2021). Science process skills in science education of developed and developing countries:

literature review. *Unnes Science Education Journal*, *10*(1), 12–17. https://doi.org/10.15294/usej.v10i1.42153
Nurdini, Y., Wulan, A. R., & Diana, S. (2020, April). Assessment for learning through written feedback to develop 21st-century critical thinking skills on plantae learning. In Journal of Physics: Conference Series (Vol. 1521, No. 4, p. 042019). https://iopscience.iop.org/article/10.1088/1742-6596/1521/4/042019/meta
Paxinou, Evgenia., Kalles, Dimitrios., Panagiotakopoulos, Christos T., Sgourou, Argyro., & Verykios, Vassilios S. (2021). An IRT-based approach to assess the learning gain of a virtual reality lab students' experience. *Intelligent Decision Technologies*, *15*(3), 487–496. https://doi.org/10.3233/IDT-200216
Pongsophon, P., & Jituafua, A. (2021). Developing and assessing learning progression for botanical literacy using Rasch analysis. *Science Education International*, *32*(2), 125–130. https://doi.org/10.33828/sei.v32.i2.5
Prasetya, A., Rosidin, U., & Herlina, K. (2019). Development of instrument assessment for learning the polytomous response models to train Higher-Order Thinking Skills (HOTS). *Journal of Physics: Conference Series*, *1155*(1). https://doi.org/10.1088/1742-6596/1155/1/012032
Prasetya, W. A., & Pratama, A. T. (2023). Item quality analysis using the Rasch model to measure critical thinking ability in the material of the human digestive system of Biology subject in high school. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *27*(1), 76–91. https://doi.org/10.21831/pep.v27i1.58873
Purnamasari, U. D., & Kartowagiran, B. (2019). Application rasch model using R program in analyze the characteristics of chemical items. *Jurnal Inovasi Pendidikan IPA*, *5*(2), 147–157. https://doi.org/10.21831/jipi.v5i2.24235
Putro, B. L., Waslaluddin, Putra, R. R. J., & Rahman, E. F. (2019). Creative learning model as implementation of curriculum 2013 to achieve 21st century skills. *Journal of Physics: Conference Series*, *1280*(3), 1–7. https://doi.org/10.1088/1742-6596/1280/3/032034
Reeve, B. B., & Masse, L. C. (2004). *Item response theory (IRT) modeling for questionnaire evaluation. In methods for testing and evaluating survey questionnaires, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer*. John Wiley & Sons.
Reise, S. P. (1982). A comparison of item-and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement Inc*, *14*(2), 127–137. https://doi.org/10.1177/014662169001400202
Retnawati, H. (2014). *Teori respon butir dan penerapannya*. www.nuhamedika.gu.ma
Retnawati, H., Munadi, S., Arlinwibowo, J., Wulandari, N. F., & Sulistyaningsih, E. (2017). Teachers' difficulties in implementing thematic teaching and learning in elementary schools. *New Educational Review*, *48*(2), 201–212. https://doi.org/10.15804/tner.2017.48.2.16
Rezba, R. J., Sprague, C. R., McDonnough, J. T., & Matkins, J. J. (1995). *Learning and assessing science process skills* (3rd ed.). Hunt Publishing Company.
Rifatul Mahmudah, I., Makiyah, Y. S., & Sulistyaningsih, D. (2019). Profil keterampilan proses sains (KPS) siswa SMA di Kota Bandung. *Diffraction*, *1*(1), 39–43. https://doi.org/10.37058/diffraction.v1i1.808
Sahin, M. G., & Yildirim, Y. (2018). The examination of item difficulty distribution, test length and sample size in different ability distribution. *Journal of Measurement and Evaluation in Education and Psychology*, *9*(3), 277–294. https://doi.org/10.21031/epod.385000
Samritin, S., & Suryanto, S. (2016). Developing an assessment instrument of junior high school students' higher order thinking skills in mathematics. *Research and Evaluation in Education*, *2*(1), 92. https://doi.org/10.21831/reid.v2i1.8268
Sarah, Wulan, A. R., & Kusnadi, M. (2020). Needs analysis on developing instruments of interpreting data skills and scientific evidence for biology learning in the 21st century era. *ACM International Conference Proceeding Series*, 188–192. https://doi.org/10.1145/3416797.3416812
Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: classical test theory vs item response theory? *Didaktika Jurnal Kependidikan*, *13*(1), 1–16. https://doi.org/10.30863/didaktika.v13i1.296
Sari, U., Duygu, E., Şen, Ö. F., & Kirindi, T. (2020). The effects of STEM education on scientific process skills and STEM awareness in simulation-based inquiry learning environment. *Journal of Turkish Science Education*, *17*(3), 387–405. https://doi.org/10.36681/tused.2020.34
Şen, C., & Vekli, G. S. (2016). The impact of inquiry-based instruction on science process skills and self-

efficacy perceptions of pre-service science teachers at a university-level biology laboratory. *Universal Journal of Educational Research*, *4*(3), 603–612. https://doi.org/10.13189/ujer.2016.040319

Sholihah, N. A. A., Sarwanto, & Aminah, N. S. (2020). Analysis of science process skill in high school students. *Journal of Physics: Conference Series*, *1567*(3). https://doi.org/10.1088/1742-6596/1567/3/032081

Siahaan, K. W. A., Lumbangaol, S. T. P., Marbun, J., Nainggolan, A. D., Ritonga, J. M., & Barus, D. P. (2020). Pengaruh model pembelajaran inkuiri terbimbing dengan multi representasi terhadap keterampilan proses sains dan penguasaan konsep IPA. *Jurnal Basicedu*, *5*(1), 195–205. https://doi.org/10.31004/basicedu.v5i1.614

Siswono, H. (2017). Analisis pengaruh keterampilan proses Sains terhadap penguasaan konsep fisika siswa. *Momentum: Physics Education Journal*, *1*(2), 83–90. https://doi.org/10.21067/mpej.v1i2.1967

Slepkov, A. D., Van Bussel, M. L., Fitze, K. M., & Burr, W. S. (2021). A baseline for multiple-choice testing in the university classroom. *SAGE Open*, *11*(2). https://doi.org/10.1177/21582440211016838

Soland, J. (2024). Item Response Theory models for difference-in-difference estimates (and whether they are worth the trouble). *Journal of Research on Educational Effectiveness*, *17*(2), 391–421. https://doi.org/10.1080/19345747.2023.2195413

Srirahayu, R. Y., & Arty, I. S. (2019). Development of experiment performance assessment instruments using guided inquiry learning models to assess science process skills. *Journal of Physics: Conference Series*, *1233*(1). https://doi.org/10.1088/1742-6596/1233/1/012075

Subali, B. (2009). Pengembangan tes pengukuran keterampilan proses sains pola divergen mata pelajaran *Biologi* SMA. *Proseding Seminar Nasional Biologi, Lingkungan Dan Pembelajarannya*, 581–593.

Supahar, S., & Prasetyo, Z. K. (2015). Pengembangan instrumen penilaian kinerja kemampuan inkuiri peserta didik pada mata pelajaran Fisika SMA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *19*(1), 96–108. https://doi.org/10.21831/pep.v19i1.4560

Thomas, M. L. (2019). Advances in applications of item response theory to clinical assessment. *Psychological Assessment*, *31*(12), 1442–1455. https://doi.org/10.1037/pas0000597

Tu, D., Gao, X., Wang, D., & Cai, Y. (2017). A new measurement of internet addiction using diagnostic classification models. *Frontiers in psychology, 8*, 1768. https://doi.org/10.3389/fpsyg.2017.01768

van der Linden, W. J., & Hambleton, R. K. (1997). Handbook of modern item response theory. In *Handbook of Modern Item Response Theory*. Springer New York. https://doi.org/10.1007/978-1-4757-2691-6

Vo, D. Van, & Simmie, G. M. (2024). Assessing scientific inquiry: A systematic literature review of tasks, tools and techniques. *International Journal of Science and Mathematics Education*, *23*, 871–906. https://doi.org/10.1007/s10763-024-10498-8

Williams-McBean, C. (2025). Factors influencing teachers' choice and use of assessment. *International Journal of Studies in Education and Science*, *6*(2), 212–239. https://doi.org/10.46328/ijses.129

Yunita, N., & Nurita, T. (2021). Analisis keterampilan proses sains siswa pada pembelajaran daring. *Pensa E-Jurnal: Pendidikan Sains*, *9*(3), 378–385. https://ejournal.unesa.ac.id/index.php/pensa

Zainul, A., & Nasoetion, , Noehi. (1997). *Penilaian hasil belajar*. Pusat Antar Universitas, Direktorat Jenderal Pendidikan Tinggi: Departemen Pendidikan Dan kebudayaan.

Zi Yan, T. G. B., & Heene, M. (2021). *Applying the Rasch model fundamental measurement in the human sciences* (4th ed.). Routledge.

Ziraluo, Y. P. B. (2021). *Pembelajaran Biologi implementasi dan pengembangan*. Forum Pemuda Aswaja.

Zuhera, Y., & Habibah, S. (2017). Kendala guru dalam memberikan penilaian terhadap sikap siswa dalam proses pembelajaran berdasarkan kurikulum 2013 di SD Negeri 14 Banda Aceh. *Jurnal Ilmiah Pendidikan Guru Sekolah Dasar*, *2*(1), 73–87. https://media.neliti.com/media/publications/187406-ID-kendala-guru-dalam-memberikan-penilaian.pdf