Prisma Sains: Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram
https://e-journal.undikma.ac.id/index.php/prismasains/index
e-mail: prismasains.pkpsm@gmail.com

October 2026. Vol. 14, No. 4
p-ISSN: 2338-4530
e-ISSN: 2540-7899
pp. 2738-2752

Preliminary Diagnostic Mapping of Pre-Service Science Teachers' Science Process Skills in General Biology: An Exploratory Rasch Analysis

[1]Ahmad Aris Arifin, [1,2]*Agus Ramdani, [1,2]A.Wahab Jufri, [3]Satutik Rahayu, [1,3]Aris Doyan [1,2]A.A Sukarso

[1]Doctoral Program of Science Education, University of Mataram, Mataram, Indonesia.
[2]Department of Biology Education, University of Mataram, Mataram, Indonesia.
[3]Department of Physics Education, University of Mataram, Mataram, Indonesia.

*Corresponding author e-mail: aramdani07@unram.ac.id

Received: August 2026; Revised: September 2026; Published: October 2026

Abstract

This preliminary diagnostic study examined the science process skills (SPS) of pre-service science teachers in General Biology and explored preliminary measurement evidence using Rasch analysis. Participants were 145 students from Biology, Chemistry, and Physics Education programs. Data were collected using 13 structured essay items representing seven basic and six integrated SPS indicators. Each indicator was represented by one item and scored using a 0–4 analytic rubric. Data were analyzed using descriptive statistics, internal-consistency analysis, and the Partial Credit Model. The mean total score was 22.959 out of 52 (44.151%), indicating that students' performance on the written SPS assessment still required improvement. Observation showed the highest raw mean (2.483), whereas operationally defining variables showed the lowest (1.241). Rasch estimates identified data interpretation as the most difficult item (0.962 logits), followed by model formulation (0.867) and hypothesis formulation (0.855), while inference was the easiest (−1.015). Several items did not elicit all upper score categories, and extreme category-level fit values indicated potential problems in category use, item demands, or differentiation within the scoring rubric. The Wright map showed that item difficulty covered a narrower range than student ability. These findings support further refinement of the items and scoring categories before the instrument is used in broader testing.

Keywords: Science process skills; Pre-service science teachers; General biology; Diagnostic assessment; Partial credit model; Rasch analysis

How to Cite: Arifin, A. A., Ramdani, A., Jufri, A. W., Rahayu, S., Doyan, A., & Sukarso, A. (2026). Preliminary Diagnostic Mapping of Pre-Service Science Teachers' Science Process Skills in General Biology: An Exploratory Rasch Analysis. *Prisma Sains: Jurnal Pengkajian Ilmu Dan Pembelajaran Matematika Dan IPA IKIP Mataram*, *14*(4), 2738–2752. https://doi.org/10.33394/j-ps.v14i4.22430

https://doi.org/10.33394/j-ps.v14i4.22430

Copyright© 2026, Arifin et al.
This is an open-access article under the CC-BY License

INTRODUCTION

Twenty-first-century science education requires learning that is oriented not only toward conceptual mastery but also toward students' ability to think scientifically, conduct investigations, evaluate evidence, and draw conclusions based on data. In this context, science process skills (SPS) are fundamental competencies because they represent the abilities required to carry out scientific work, such as observing, classifying, measuring, predicting, interpreting data, formulating hypotheses, designing experiments, and communicating results. SPS also bridge conceptual understanding and scientific practice; therefore, science learning should provide students with opportunities to experience the process of acquiring knowledge rather than merely receiving scientific knowledge. Numerous studies have shown that SPS can be strengthened through inquiry-based learning, laboratory work, experimentation, discovery learning, projects, and other active-learning strategies (Astalini et al., 2023; Bahtiar & Dukomalamo, 2019; Idris et al., 2022; Nunaki et al., 2020; Suryanti et al., 2020; Wardani & Djukri, 2019).

SPS occupy an important position in the education of pre-service science teachers. Pre-service teachers are expected not only to understand scientific concepts but also to master

scientific ways of working so that they can later design learning based on investigation, laboratory activities, and problem solving. Several studies involving pre-service science teachers have indicated that SPS still need to be strengthened systematically through laboratory work and inquiry-oriented learning (Adlim et al., 2018; Artayasa et al., 2023; Hacıeminoğlu et al., 2022; Harahap et al., 2019; Hiğde & Aktamış, 2022; Jannah et al., 2025; Wola et al., 2023). In the context of General Biology, SPS are particularly relevant because the course covers biological phenomena that can be studied through observation, measurement, classification, data interpretation, and inference. Previous studies have also shown that SPS can be profiled through General Biology and laboratory activities (Susanti & Anwar, 2018; Wola et al., 2023).

Although SPS have been widely investigated, many previous studies have focused on the effects of instructional models on SPS improvement, including discovery learning, guided inquiry, open inquiry, project-based learning, STEM/STEAM, and blended learning (Bahtiar & Dukomalamo, 2019; Harahap et al., 2019; Hiğde & Aktamış, 2022; Indriani & Mercuriani, 2019; Mufida et al., 2020; Rusmini et al., 2021; Wardani & Djukri, 2019). These studies demonstrate that SPS can be developed through instructional interventions. However, before an intervention is designed, an initial diagnostic mapping is needed to identify patterns of students' SPS, including areas of relative strength and weakness, as well as the characteristics of the assessment items used to measure them. Such information is important as an empirical basis for refining the assessment and planning subsequent instruction.

One limitation of SPS assessment based solely on total scores, means, and percentages is that these statistics provide an overall description of achievement but do not directly explain the relationship between respondent ability and item difficulty, item fit, or the functioning of response categories. Rasch analysis can complement descriptive statistics by placing respondent ability and item difficulty on the same logit scale and by providing information about item fit, response-category functioning, and person–item targeting (Boone, 2016; Planinic et al., 2019). Rasch analysis has also been applied in SPS and science-education research for instrument evaluation and diagnostic purposes (Handayani, 2021; Hiğde et al., 2024; Mustafa et al., 2021; Rahayu et al., 2021). Therefore, the use of Rasch analysis in the present study is not considered a methodological novelty, but rather an exploratory approach that complements raw-score information in examining students' initial SPS and the functioning of the assessment.

The novelty of this study lies in its focus on the preliminary diagnostic mapping of pre-service science teachers' SPS in General Biology by integrating raw-score analysis with exploratory Rasch evidence. This integration enables the initial SPS profile to be examined in relation to item difficulty, response-category functioning, and the distribution of student ability, while also identifying aspects of the assessment items and scoring rubric that require refinement. Unlike studies that primarily examine the effectiveness of instructional interventions, the present study positions diagnostic baseline mapping as an empirical foundation for instrument refinement and subsequent instructional planning.

Based on this background, this study aimed to describe the initial SPS of pre-service science teachers in General Biology and to examine preliminary measurement evidence using the Partial Credit Model. Specifically, the study examined SPS achievement, item difficulty and fit, response-category functioning, local dependence, and person–item targeting to identify aspects of the assessment that require refinement and to provide baseline information for subsequent instructional design.

METHOD

Research Design

This study employed a quantitative approach with a cross-sectional diagnostic design. It was conducted as a preliminary study to map the initial science process skills of pre-service science teachers in the context of a General Biology course. The mapping was undertaken before the implementation of any specific instructional model or laboratory intervention. The study was not intended to test causal relationships or the effectiveness of a treatment; rather, it

aimed to generate an empirical profile of mastery for each science process skill indicator, the distribution of student ability, item difficulty, and the psychometric functioning of the instrument through exploratory Rasch analysis.

Research Participants

The study involved 145 pre-service science teachers enrolled in the Biology Education, Chemistry Education, and Physics Education programs at Universitas Pendidikan Mandalika. The participants comprised 89 Biology Education students, 31 Chemistry Education students, and 25 Physics Education students. By semester level, 54 students were in the second semester, 58 in the fourth semester, one in the sixth semester, and 32 in the eighth semester.

Participants were selected using purposive sampling based on the following criteria: they were active students in one of the science education programs and had completed or were currently taking General Biology. Prior experience with General Biology was required to ensure that the students possessed sufficient content knowledge to understand the item stimuli. Differences in the numbers of students across academic programs and semesters meant that group analyses were used for descriptive diagnostic mapping rather than for drawing inferential conclusions about between-group differences.

Research Instrument and Instrument Development

Data were collected using a science process skills (SPS) test developed in the context of General Biology. The instrument consisted of 13 structured essay items representing 13 SPS indicators. Seven indicators represented basic SPS: observation, classification, spatial and temporal relationships, using numbers, measurement, inference, and prediction. The remaining six represented integrated SPS: operationally defining variables, formulating models, controlling variables, interpreting data, formulating hypotheses, and designing experiments. Each SPS indicator was represented by one item, resulting in a total of 13 items with a maximum overall score of 52.

Item development began by identifying the SPS indicators to be assessed. Each indicator was then translated into one written task using a common stimulus context, namely seedling growth. The stimulus and questions were designed to elicit students' scientific reasoning and procedural understanding in accordance with the SPS indicator being assessed. For indicators that are commonly assessed through direct laboratory performance, the written items focused on the reasoning and procedural-planning components underlying the corresponding scientific processes. Accordingly, the instrument was used to obtain an initial profile of students' SPS based on written responses and was not intended to replace direct assessment of laboratory performance.

The instrument did not undergo expert-based content validation prior to administration. The internal consistency of the instrument scores was subsequently examined using Cronbach's alpha. This analysis was used to evaluate the consistency among the items. The analysis yielded a Cronbach's alpha coefficient of 0.862, indicating good internal consistency.

Scoring Rubric and Scoring Procedure

Each item was scored using a 0–4 analytic rubric developed on the basis of the expected response components for the corresponding SPS indicator. The score categories reflected increasing levels of completeness, accuracy, and appropriateness of students' responses, while the descriptor for each category served as the scoring criterion. Because each SPS indicator was represented by one item, the maximum score for each indicator was 4. All student responses were scored by a single rater using the same analytic rubric. A single rater was used to maintain consistency in the application of the scoring criteria across all student responses.

Data Collection Procedures

Data were collected in May 2026 during the even semester of the 2025–2026 academic year. Before the test was administered, students received an explanation of the study objectives, test procedures, data use, and the confidentiality of participant identities. The test was provided in printed form and completed individually within a predetermined period. Students were

instructed to read the stimuli carefully and provide clear, systematic responses. After completion, their responses were collected and scored using the established rubric. Scores were entered into a data matrix containing participant codes, academic programs, semesters, and scores for KPS1 through KPS13. Students' personal identities were removed or replaced with codes before analysis. The final dataset comprised 145 participants and 13 polytomously scored items ranging from 0 to 4.

Data Analysis

Data analysis was conducted using R software version [R version] with the psych, TAM, and WrightMap packages. The analysis proceeded in stages, including data screening, descriptive analysis, internal-consistency testing, Rasch analysis using the Partial Credit Model, examination of score-category functioning, item-fit analysis, assessment of local dependence, and construction of the Wright map. The analytical procedures produced descriptive statistics, summaries by academic program and semester, Cronbach's alpha coefficients, item-difficulty parameters, item-fit statistics, participant-ability estimates, and a Wright map.

Data Screening and Preparation

Before the main analyses were conducted, the data were examined to ensure completeness and scoring accuracy. Screening included identifying duplicate records, missing responses, coding errors, and scores outside the allowable range of 0–4. The frequency distribution of categories for each item was also examined to determine which score categories were used frequently, used rarely, or not used by students.

Each row of the dataset represented one student, while columns KPS1 through KPS13 represented scores on the respective indicators. The total score was calculated by summing the scores across all items, and the percentage score was obtained by dividing the total score by the maximum score of 52. Missing responses, when present, were not automatically assigned a score of zero. Such responses were retained as missing data in Rasch estimation, whereas conventional total scores were calculated only for data meeting the specified completeness criteria.

Descriptive Analysis

Descriptive analysis was used to portray students' initial science process skills. Statistics calculated for each item included the number of valid responses, mean, standard deviation, minimum score, and maximum score. Total scores, percentage scores, score ranges, and the distribution of student ability were also calculated.

The mean score for each item was used to identify SPS indicators that had been relatively mastered and those that still showed low achievement. A higher mean indicated that students were better able to respond in accordance with the criteria for the indicator, whereas a lower mean indicated greater difficulty. This descriptive interpretation was distinguished from item difficulty in the Rasch analysis because raw-score means and logit difficulty parameters do not necessarily produce identical rankings.

Descriptive analyses were also conducted by academic program and semester. Mean total scores and percentages were calculated for Biology Education, Chemistry Education, and Physics Education, as well as for each semester level. These analyses were used only to describe patterns within the study sample. Inferential tests of group differences were not conducted because group sizes were unequal, particularly for the sixth semester, which included only one participant.

Internal-Consistency Analysis

The internal consistency of the instrument was analyzed using Cronbach's alpha with the psych package. The indices examined included raw alpha, standardized alpha, the average inter-item correlation, and Guttman's lambda-6. Higher coefficient values indicate greater consistency among the items in measuring the science process skills construct.

The primary analysis used the original score direction because all items were designed positively, meaning that higher scores indicated higher SPS ability. The check.keys = TRUE

option was used as an additional check to detect any item that might empirically function in the opposite direction. However, a decision to reverse-score an item was not based solely on software output; it also considered the item content, the indicator being measured, and the scoring rubric. The alpha-if-item-deleted statistic was also examined to identify any item that might reduce the instrument's internal consistency.

Cronbach's alpha was used as evidence of the consistency of raw scores, but it was not treated as the sole indicator of instrument quality. The results were interpreted together with the reliability of ability estimates, item-fit statistics, score-category functioning, and the distribution of items in the Rasch analysis.

Exploratory Rasch Analysis

Exploratory Rasch analysis was conducted using the Partial Credit Model (PCM) through the TAM package. The PCM was selected because each item was assigned ordered polytomous scores from 0 to 4. Unlike a dichotomous model, the PCM estimates the probability of students attaining each score category and allows the category-threshold structure to vary across items.

The model was estimated using the tam.mml() function with the specification irtmodel = "PCM2". Estimation was performed using marginal maximum likelihood. This analysis placed student ability and item difficulty on the same measurement scale, namely the logit scale. Because the study was preliminary, Rasch analysis was used to obtain diagnostic information about patterns of student ability and item characteristics rather than to claim that the instrument had undergone final validation.

Item-Difficulty Analysis

The difficulty of each item was determined from the item-location parameter, or xsi.item. Higher positive logit values indicate that an item or SPS indicator is relatively more difficult for students to attain. Conversely, lower or negative logit values indicate that an item is relatively easier.

Items were then ordered from the highest to the lowest logit value to identify the most challenging and easiest SPS indicators. Difficulty was interpreted not only from the logit value but also by considering the task format and indicator complexity. For example, interpreting data, formulating hypotheses, or designing experiments may require the integration of several cognitive processes compared with indicators focused on identifying facts.

Person-Ability Estimation

Each student's ability was estimated using the expected a posteriori (EAP) value and expressed as theta logits. A higher theta value indicates stronger initial science process skills, whereas a lower theta value indicates that a student experienced difficulty meeting the demands of most items.

Rasch ability estimates were used to complement raw total scores. Total scores treat each score increment as equally spaced, whereas Rasch estimates account for item difficulty and students' response patterns. EAP reliability was also calculated to indicate the consistency with which students' abilities could be differentiated based on their responses to the instrument.

Response-Category Functioning

Score-category functioning was examined to determine whether categories 0, 1, 2, 3, and 4 operated in accordance with the intended order of mastery. The analysis included category-use frequencies, category thresholds, and the ordering of thresholds. Categories with very low frequencies or no observations may produce unstable estimates and extreme outfit statistics.

Preliminary results indicated that not all items empirically used the full score range. Therefore, rarely used categories were not immediately collapsed or removed. Instead, they were first examined in relation to rubric clarity, possible overlap among criteria, the quality of students' responses, and the alignment between item demands and score categories. This information was used as a basis for revising the rubric in subsequent research.

Item-Fit Analysis

Item fit to the Rasch model was examined using infit mean-square and outfit mean-square statistics. Values close to 1.00 indicate that students' response patterns are consistent with model expectations. As an exploratory guideline, mean-square values between 0.50 and 1.50 were regarded as productive for measurement. Values above 1.50 indicate greater-than-expected response variation, or underfit, whereas values below 0.50 indicate overly predictable responses, or overfit.

Infit statistics are more sensitive to unexpected responses on items whose difficulty is close to a student's ability, whereas outfit statistics are more sensitive to extreme or unexpected responses on items far from a student's ability level. Standardized statistics and probability values were used as supplementary information but were not treated as the sole basis for determining item adequacy.

For the main presentation in the article, fit statistics were summarized at the item level, with one row for each SPS indicator. Category-level fit statistics were used as additional diagnostic information and may be placed in an appendix or supplementary material. Items showing misfit were not automatically removed; rather, they were examined in terms of item wording, stimulus clarity, scoring-rubric quality, category frequencies, and alignment with the SPS indicator.

Local-Independence Analysis

Local dependence among items was examined through residual correlations using the Q3 or adjusted-Q3 statistic. The local-independence principle assumes that, after students' SPS ability has been controlled, the response to one item should not be strongly influenced by responses to other items.

Item pairs with relatively high positive residual correlations were examined to identify possible similarities in content, repeated response demands, or dependence arising because all items used the same seedling-growth stimulus. Because the instrument was designed around a shared stimulus, the Q3 results were interpreted exploratorily and substantively. High residual correlations were not used as an automatic basis for deleting items without considering the representation of SPS indicators.

Wright Map and Test Targeting

The Wright map was used to display the distribution of student ability and item difficulty on the same logit scale. The distribution of student ability was placed on one side of the map, while the locations of items or SPS indicators were placed on the other. Students positioned higher on the map had relatively higher SPS ability, whereas items positioned higher represented relatively more difficult indicators. The Wright map was used to evaluate instrument targeting, namely the correspondence between the range of item difficulty and the range of student ability. If most items are located far below students' abilities, the instrument tends to be too easy. Conversely, if most items are located far above students' abilities, the instrument tends to be too difficult. Ability regions without corresponding items indicate measurement gaps and can provide a basis for developing new items. In addition, the position of each indicator on the Wright map was used to construct a diagnostic SPS profile. Indicators located above the mean difficulty were regarded as more challenging skills, whereas those located below the mean were relatively easier for students to attain.

Diagnostic Mapping

Diagnostic mapping was conducted by integrating descriptive scores, item difficulty, student-ability estimates, item-fit statistics, category functioning, and the Wright map. Indicators were classified not only according to high or low mean scores but also according to their positions on the logit scale and their fit to the model.

The mapping results were used to identify three main components: SPS indicators that had been relatively mastered, indicators still requiring strengthening, and indicators requiring review

of the instrument or rubric. This mapping provides a basis for designing more responsive General Biology instruction, including the selection of laboratory activities, investigative tasks, data-interpretation exercises, hypothesis formulation, and experimental design.

Interpretation of the Exploratory Results

All Rasch results were interpreted as exploratory information because this was a preliminary study and each indicator was represented by only one item. Item misfit, suboptimal category functioning, or gaps in the Wright map were not used directly to conclude that an indicator was invalid. Instead, these findings were treated as diagnostic evidence for refining stimuli, item wording, scoring rubrics, and the development of instruments containing multiple items for each indicator.

RESULTS AND DISCUSSION

This section presents the results of the diagnostic mapping of the initial science process skills of pre-service science teachers in the General Biology course. The analysis focused on identifying participants' overall ability levels, the distribution of mastery across the measured science process skill indicators, and instrument-item characteristics based on exploratory Rasch analysis. The findings are discussed by integrating descriptive analysis, measures of respondent ability and item difficulty, reliability and separation indices, item fit, and the Wright map.

Table 1. General Summary of the SPS Data

Component	Value
Number of respondents	145
Number of SPS items	13
Item score range	0–4
Mean total score	22.959
Mean percentage	44.151%
Minimum total score	5
Maximum total score	31

Table 2. Descriptive Statistics for Each SPS Item

Indicator	Mean	SD	Min	Max
KPS1	2.483	0.826	0	4
KPS2	2.317	0.761	0	4
KPS3	2.200	0.760	0	4
KPS4	1.283	0.752	0	2
KPS5	1.959	0.881	0	4
KPS6	2.083	0.672	0	3
KPS7	1.800	0.703	0	4
KPS8	1.241	0.739	0	2
KPS9	1.400	0.776	0	3
KPS10	1.717	0.779	0	4
KPS11	1.517	0.809	0	4
KPS12	1.359	0.822	0	4
KPS13	1.600	0.837	0	4

Based on the item means, the highest achievement was found for KPS1 (M = 2.483), whereas the lowest was found for KPS8 (M = 1.241). A lower mean value indicates lower student achievement on the corresponding item.

Table 3. SPS Results by Semester

Semester	Total score	Percentage
2	18.852	36.254
4	25.759	49.536
6	7.000	13.462
8	25.312	48.678

Table 4. Rasch Item Parameters Based on the R Output

Item	N	M	xsi.item	AXsi_.Cat1	AXsi_.Cat2	AXsi_.Cat3	AXsi_.Cat4	B.Cat1.Dim1	B.Cat2.Dim1	B.Cat3.Dim1	B.Cat4.Dim1
KPS1	145	2.483	-0.451	-2.005	-4.298	-4.635	-1.803	1	2	3	4
KPS2	145	2.317	-0.202	-0.967	-4.439	-3.926	-0.809	1	2	3	4
KPS3	145	2.200	-0.068	0.093	-4.161	-3.095	-0.270	1	2	3	4
KPS4	145	1.283	-0.595	-1.127	-1.190	—	—	1	2	0	0
KPS5	145	1.959	0.684	1.998	-2.698	-1.294	2.737	1	2	3	4
KPS6	145	2.083	-1.015	-1.649	-4.380	-3.044	—	1	2	3	0
KPS7	145	1.800	0.491	-0.035	-3.148	0.896	1.965	1	2	3	4
KPS8	145	1.241	-0.528	-1.214	-1.057	—	—	1	2	0	0
KPS9	145	1.400	0.867	-1.050	-1.553	2.601	—	1	2	3	0
KPS10	145	1.717	0.515	-0.166	-2.622	2.487	2.061	1	2	3	4
KPS11	145	1.517	0.962	-0.670	-1.847	2.435	3.847	1	2	3	4
KPS12	145	1.359	0.855	-1.228	-1.323	2.671	3.422	1	2	3	4
KPS13	145	1.600	0.637	-0.462	-2.066	2.955	2.549	1	2	3	4

Note. "—" indicates that the category parameter was not estimated because the corresponding score category did not appear in the response data.

Table 5. Item-Difficulty Ranking Based on xsi.item Values

Item	M	Xsi.item
KPS11	1.517	0.962
KPS9	1.400	0.867
KPS12	1.359	0.855
KPS5	1.959	0.684
KPS13	1.600	0.637
KPS10	1.717	0.515
KPS7	1.800	0.491
KPS3	2.200	-0.068
KPS2	2.317	-0.202
KPS1	2.483	-0.451
KPS8	1.241	-0.528
KPS4	1.283	-0.595
KPS6	2.083	-1.015

Interpretation of xsi.item values: a higher logit value indicates that an item is relatively more difficult, whereas a lower logit value indicates that an item is relatively easier for respondents to answer or attain.

Table 6. Score-Category Functioning Across SPS Items

Item	0	1	2	3	4	Missing categories	Thresholds (δ1–δ4)	Ordering
KPS1	5	9	49	75	7	None	-2.005; -2.293; -0.337; 2.832	Disordered
KPS2	6	4	77	54	4	None	-0.967; -3.472; 0.512; 3.118	Disordered
KPS3	8	2	92	39	4	None	0.093; -4.254; 1.066; 2.825	Disordered
KPS4	26	52	67	0	0	3, 4	-1.127; -0.063	Ordered
KPS5	19	1	93	31	1	None	1.998; -4.697; 1.404; 4.031	Disordered
KPS6	6	9	97	33	0	4	-1.649; -2.731; 1.336	Disordered
KPS7	15	6	119	3	2	None	-0.035; -3.113; 4.044; 1.069	Disordered
KPS8	26	58	61	0	0	3, 4	-1.214; 0.157	Ordered
KPS9	24	41	78	2	0	4	-1.050; -0.503; 4.154	Ordered
KPS10	19	10	112	1	3	None	-0.166; -2.457; 5.109; -0.426	Disordered
KPS11	24	26	92	2	1	None	-0.670; -1.177; 4.282; 1.413	Disordered
KPS12	24	51	66	2	2	None	-1.228; -0.096; 3.994; 0.751	Disordered
KPS13	23	19	99	1	3	None	-0.462; -1.604; 5.021; -0.406	Disordered

Note. Category frequencies are based on 145 responses per item. Thresholds were derived from adjacent PCM category intercepts; only empirically observed categories were included. "Disordered" indicates that the estimated step thresholds did not increase monotonically. Missing upper categories occurred for KPS4 (scores 3–4), KPS6 (score 4), KPS8 (scores 3–4), and KPS9 (score 4).

Table 7. Category-Level Rasch Fit Results Based on the R Output

Parameter	Outfit	Outfit_t	Outfit_p	Outfit_pholm	Infit	nfit_t	Infit_p	Infit_pholm
KPS1_Cat1	0.213	-3.714	2.037e-04	0.004	1.003	0.084	0.933	1.000
KPS1_Cat2	0.290	-4.795	1.626e-06	4.878e-05	0.756	-1.215	0.224	1.000
KPS1_Cat3	1.284	3.551	3.842e-04	0.008	1.179	2.343	0.019	0.726
KPS1_Cat4	21.033	14.646	1.437e-48	5.890e-47	1.137	0.501	0.616	1.000
KPS2_Cat1	32.162	17.676	6.443e-70	2.771e-68	1.219	0.782	0.434	1.000
KPS2_Cat2	2.857	4.452	8.508e-06	2.297e-04	1.057	0.295	0.768	1.000
KPS2_Cat3	1.928	13.113	2.788e-39	1.115e-37	1.276	4.595	4.330e-06	1.992e-04
KPS2_Cat4	60.603	18.377	1.990e-75	8.757e-74	1.097	0.352	0.725	1.000
KPS3_Cat1	1.573	0.827	0.408	1.000	1.175	0.724	0.469	1.000
KPS3_Cat2	0.762	-1.400	0.161	1.000	1.078	0.372	0.710	1.000
KPS3_Cat3	3.739	21.697	2.213e-104	9.958e-103	1.261	3.211	0.001	0.056
KPS3_Cat4	152.150	26.182	4.307e-151	1.981e-149	1.064	0.283	0.777	1.000
KPS4_Cat1	0.703	-2.360	0.018	0.329	0.586	-3.100	0.002	0.079
KPS4_Cat2	0.915	-1.518	0.129	1.000	1.006	0.111	0.912	1.000
KPS5_Cat1	5.171	8.118	4.731e-16	1.798e-14	0.944	-0.205	0.837	1.000
KPS5_Cat2	3.118	5.329	9.857e-08	3.154e-06	0.845	-0.706	0.480	1.000
KPS5_Cat3	2.214	8.502	1.866e-17	7.279e-16	1.129	1.228	0.219	1.000
KPS5_Cat4	0.409	-0.570	0.568	1.000	0.974	0.291	0.771	1.000
KPS6_Cat1	0.204	-4.073	4.632e-05	0.001	0.970	-0.026	0.980	1.000
KPS6_Cat2	0.401	-3.824	1.313e-04	0.003	0.749	-1.280	0.200	1.000
KPS6_Cat3	1.179	1.630	0.103	1.000	1.038	0.389	0.697	1.000
KPS7_Cat1	0.314	-4.547	5.438e-06	1.523e-04	0.851	-0.690	0.490	1.000
KPS7_Cat2	0.784	-1.338	0.181	1.000	0.668	-1.879	0.060	1.000
KPS7_Cat3	8.094	7.042	1.895e-12	7.013e-11	1.047	0.241	0.809	1.000
KPS7_Cat4	137.717	16.244	2.481e-59	1.042e-57	0.988	0.189	0.850	1.000
KPS8_Cat1	1.132	0.548	0.584	1.000	0.641	-2.640	0.008	0.332
KPS8_Cat2	0.932	-1.260	0.208	1.000	1.030	0.528	0.597	1.000
KPS9_Cat1	1.109	0.178	0.859	1.000	0.617	-2.636	0.008	0.332
KPS9_Cat2	1.077	1.042	0.297	1.000	1.082	1.128	0.259	1.000
KPS9_Cat3	9.392	4.669	3.024e-06	8.770e-05	1.001	0.224	0.822	1.000
KPS10_Cat1	0.271	-5.232	1.676e-07	5.195e-06	0.624	-2.117	0.034	1.000
KPS10_Cat2	0.862	-0.845	0.398	1.000	0.849	-0.897	0.370	1.000
KPS10_Cat3	3.279	2.953	0.003	0.060	0.998	0.131	0.895	1.000
KPS10_Cat4	6.492	3.831	1.275e-04	0.003	0.972	0.103	0.918	1.000
KPS11_Cat1	0.232	-6.580	4.709e-11	1.695e-09	0.429	-4.174	2.993e-05	0.001
KPS11_Cat2	0.932	-0.662	0.508	1.000	0.924	-0.725	0.468	1.000
KPS11_Cat3	1.219	0.416	0.677	1.000	0.964	0.103	0.918	1.000
KPS11_Cat4	0.407	-0.655	0.513	1.000	0.976	0.284	0.777	1.000
KPS12_Cat1	0.266	-6.561	5.333e-11	1.867e-09	0.468	-4.097	4.180e-05	0.002
KPS12_Cat2	0.940	-1.017	0.309	1.000	1.024	0.397	0.691	1.000
KPS12_Cat3	3.769	3.244	0.001	0.024	0.995	0.125	0.900	1.000
KPS12_Cat4	2.945	1.506	0.132	1.000	0.987	0.178	0.859	1.000
KPS13_Cat1	0.239	-6.187	6.115e-10	2.079e-08	0.469	-3.607	3.095e-04	0.013
KPS13_Cat2	0.891	-0.887	0.375	1.000	0.870	-1.046	0.295	1.000
KPS13_Cat3	4.614	4.073	4.650e-05	0.001	1.045	0.236	0.813	1.000
KPS13_Cat4	10.502	5.648	1.622e-08	5.353e-07	1.030	0.215	0.830	1.000

Table 7 presents category-level fit statistics. Several extreme outfit values occurred in sparsely used upper categories, indicating that category misfit should not be attributed solely to students' ability. The pattern may also reflect limited category use, item-specific response demands, or insufficient differentiation among the 0–4 rubric descriptors.

Table 8. Item-Level Rasch Fit Statistics for the SPS Items

Item	Infit MNSQ	Outfit MNSQ	Interpretation
KPS1	1.145	0.968	Productive fit
KPS2	1.465	1.247	Productive fit
KPS3	1.565	1.071	Underfit
KPS4	0.676	0.724	Productive fit
KPS5	1.098	0.931	Productive fit
KPS6	0.800	0.744	Productive fit
KPS7	0.852	0.432	Overfit
KPS8	0.748	0.844	Productive fit
KPS9	0.845	0.895	Productive fit
KPS10	0.762	0.517	Productive fit
KPS11	0.590	0.567	Productive fit
KPS12	0.698	0.664	Productive fit
KPS13	0.691	0.539	Productive fit

Note. Item-level mean-square fit was calculated from observed scores and PCM model-expected scores using the available EAP ability estimates. As an exploratory guideline, MNSQ values from 0.50 to 1.50 were treated as productive for measurement.

At the item level, most SPS items showed productive fit to the PCM. KPS3 showed slight underfit (Infit MNSQ = 1.565), whereas KPS7 showed overfit in the outfit statistic (Outfit MNSQ = 0.432). These results were interpreted together with category frequencies and category-level fit rather than as an automatic basis for item deletion.

Table 9. Item Pairs with the Highest Q3 Residual Correlations

Item Pair	SPS Indicators	Q3 Residual Correlation	Adjusted Q3	Interpretation
KPS10–KPS13	Controlling Variables – Designing Experiments	0.397	0.473	Possible local dependence
KPS10–KPS11	Controlling Variables – Interpreting Data	0.388	0.464	Possible local dependence
KPS11–KPS12	Interpreting Data – Formulating Hypotheses	0.375	0.451	Possible local dependence
KPS11–KPS13	Interpreting Data – Designing Experiments	0.329	0.405	Possible local dependence
KPS12–KPS13	Formulating Hypotheses – Designing Experiments	0.269	0.346	Possible local dependence
KPS10–KPS12	Controlling Variables – Formulating Hypotheses	0.216	0.292	Possible local dependence
KPS1–KPS5	Observation – Measurement	0.147	0.223	Possible local dependence
KPS6–KPS7	Inference – Prediction	0.143	0.219	Possible local dependence
KPS1–KPS2	Observation – Classification	0.129	0.205	Possible local dependence

Local dependence was examined using Yen's Q3 residual correlations. The mean residual correlation across item pairs was −0.076. After adjustment for the mean residual correlation, several item pairs showed adjusted Q3 values above 0.20, suggesting possible local dependence. The strongest residual associations were observed for KPS10–KPS13 (adjusted Q3 = 0.473), KPS10–KPS11 (0.464), KPS11–KPS12 (0.451), and KPS11–KPS13 (0.405). These relatively high residual correlations were mainly concentrated among the integrated SPS items.

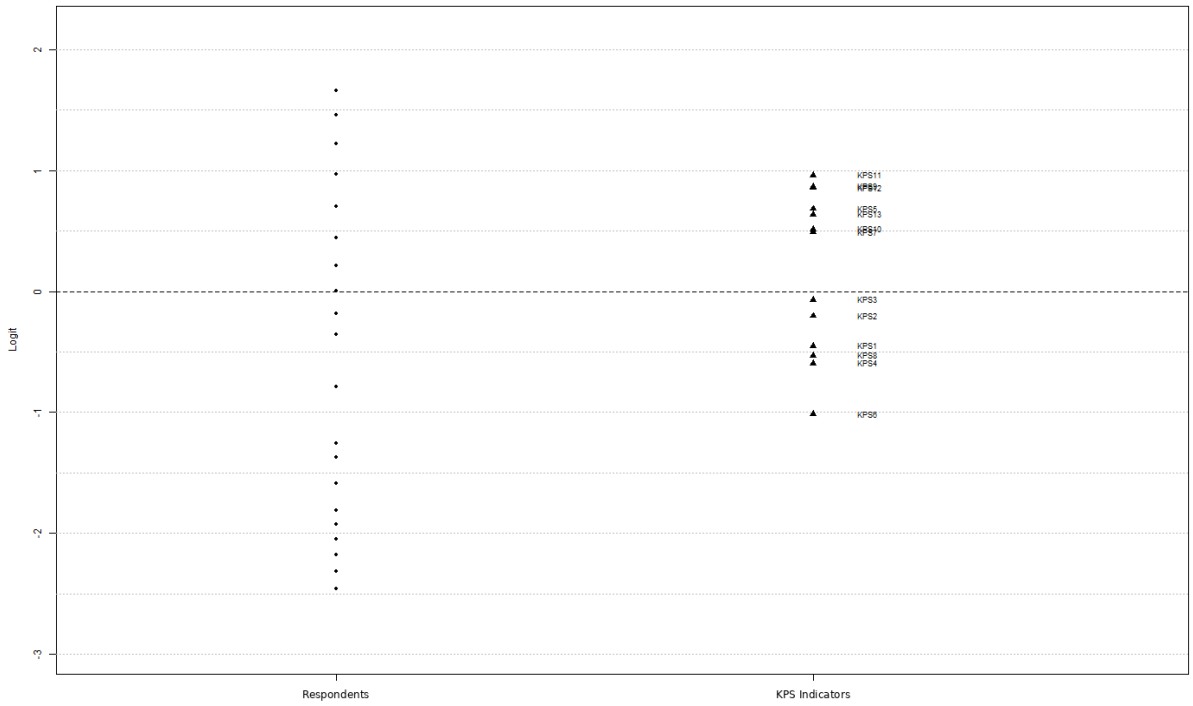


Figure 1. SPS Wright Map with Reference Lines (Logit Scale)

The Wright map shows respondent ability and SPS item difficulty on the same logit scale. Respondent ability covered a wider range than item difficulty. SPS11 was the most difficult item (0.962 logits), whereas SPS6 was the easiest (−1.015 logits). The narrower item range indicates incomplete person–item targeting and measurement gaps at the upper and lower ends of the ability continuum.

Discussion

Overall, the written assessment results indicate that pre-service science teachers experienced difficulty meeting several response demands across the SPS indicators. The mean total score was 22.959 out of 52 (44.151%). This pattern is consistent with previous reports showing variation in SPS achievement among pre-service teachers and the need for continued opportunities to engage in scientific investigation (Susanti & Anwar, 2018; Wola et al., 2023). Idris et al. (2022) likewise emphasized the role of meaningful experimental activity and active engagement in developing SPS. Because the present instrument was a stimulus-based written test, however, the scores primarily represent students' written scientific reasoning and procedural planning rather than their actual performance in laboratory settings.

The raw-score pattern suggests that several basic SPS items were answered more successfully than several integrated SPS items in this assessment. Observation (SPS1) showed the highest mean (M = 2.483), followed by classification (SPS2; M = 2.317) and spatial and temporal relationships (SPS3; M = 2.200). Lower means were observed for several integrated items, including operationally defining variables (SPS8; M = 1.241), formulating hypotheses (SPS12; M = 1.359), formulating models (SPS9; M = 1.400), interpreting data (SPS11; M = 1.517), and designing experiments (SPS13; M = 1.600). Because each indicator was represented by only one item, this pattern should be interpreted at the item level rather than as definitive evidence that integrated SPS are generally weaker than basic SPS. The differences may reflect both the reasoning demands of the indicators and the specific wording and context of the items.

The raw-score ranking was not fully consistent with the Rasch difficulty hierarchy. SPS8 had the lowest raw mean, yet its item location was -0.528 logits and it was not the most difficult Rasch item. In addition, responses to SPS8 used only score categories 0-2, so categories 3 and 4 could not be estimated. Similar restrictions occurred for SPS4, SPS6, and SPS9. The discrepancy between the raw mean and Rasch estimate for SPS8 illustrates measurement

uncertainty when interpretation depends on a single item with a restricted empirical score range. Its low raw mean may therefore reflect a combination of student responses, item demands, and limited differentiation among rubric categories rather than difficulty alone (Boone, 2016; Planinic et al., 2019).

Within this written assessment, the Rasch difficulty hierarchy placed data interpretation (SPS11; 0.962 logits) as the most challenging item, followed by model formulation (SPS9; 0.867 logits) and hypothesis formulation (SPS12; 0.855 logits). These tasks required students to identify patterns, explain relationships, construct representations of variables, and formulate testable propositions. Inference (SPS6; -1.015 logits) was relatively the easiest item. The observed hierarchy therefore reflects the interaction between the reasoning required by each task and the way that task was represented in the written stimulus. It should not be interpreted as a direct ranking of students' laboratory performance across SPS domains (Mustafa et al., 2021; Hacıeminoğlu et al., 2022).

The measurement item (SPS5) also illustrates the distinction between written SPS reasoning and direct laboratory performance. Although measurement is categorized as a basic SPS, SPS5 was relatively difficult (0.684 logits) because students were asked to identify the appropriate instrument, unit, object component, and measurement procedure in writing. The result therefore indicates difficulty with written procedural reasoning rather than direct evidence that students are unable to perform measurements in the laboratory. Future assessment work should combine written tasks with performance-based observations so that procedural reasoning and actual laboratory performance can be examined together.

Scores by academic program and semester are presented as descriptive characteristics of the sample rather than as evidence of group differences. Chemistry Education, Physics Education, and Biology Education showed mean percentages of 49.752%, 49.077%, and 40.817%, respectively, while semester-level percentages also varied. These patterns should not be interpreted comparatively because the group sizes were unequal and the sixth semester was represented by only one participant. More balanced samples would be required for meaningful comparisons across programs or semester levels.

The instrument showed good internal consistency (Cronbach's alpha = 0.862), but several measurement features indicate that the results should remain exploratory. Each SPS indicator was represented by one item, several upper score categories were rarely used or absent, and category-level fit statistics included extreme outfit values. In addition, all items were based on the same seedling-growth stimulus. The Q3 results showed relatively high residual associations among several item pairs, especially within the integrated SPS items, suggesting possible local dependence. A misunderstanding of the shared stimulus could therefore influence responses to more than one item at the same time. For this reason, the observed SPS profile may partly reflect the specific seedling-growth context and should not be generalized directly to other biological contexts or to SPS performance more broadly (Boone, 2016; Rahayu et al., 2021; Hiğde et al., 2024).

The Wright map further indicates that the item set did not fully target the range of respondent ability. Item locations ranged from approximately -1.015 to 0.962 logits, whereas respondent ability extended beyond this range at both ends of the continuum. The instrument therefore provided less information for respondents with very low or very high estimated ability. Future development should add multiple items for each indicator and distribute them across a wider range of difficulty so that the instrument can better represent the ability continuum (Planinic et al., 2019; Mustafa et al., 2021).

The diagnostic pattern can inform the design of future instructional and intervention studies. In particular, subsequent work may examine activities that progressively engage students in interpreting data, defining and controlling variables, formulating models and hypotheses, and designing investigations. Inquiry-oriented, open-inquiry, project-based, and blended-learning approaches have been associated with SPS development in previous studies (Artayasa et al., 2023;

Harahap et al., 2019; Jannah et al., 2025; Mufida et al., 2020; Rusmini et al., 2021). In the present context, these approaches are theoretically grounded directions for future intervention research rather than instructional effects demonstrated by this diagnostic study.

These findings should be interpreted within the scope of a preliminary, context-specific assessment. The use of one shared seedling-growth stimulus may have contributed to local dependence, one item per indicator limited construct representation, and the written format did not directly assess procedural execution in the laboratory. Purposive sampling and unequal group sizes also limit generalizability. Further instrument development should use multiple biological contexts, include several items for each indicator, refine score-category descriptors, involve more than one rater, and combine written assessment with laboratory performance measures. These steps would allow later studies to examine the stability of the SPS profile across contexts and provide a stronger basis for subsequent intervention research.

Limitations

This study has several limitations that should be considered when interpreting the findings. Each SPS indicator was represented by only one item, which limited the breadth of construct representation and made indicator-level interpretations dependent on the functioning of a single task. In addition, all items were developed using the same shared stimulus on seedling growth and development. Although this provided a consistent context across items, misunderstanding the shared stimulus may have influenced responses to several items simultaneously, and the residual-correlation results also indicated possible local dependence among several item pairs, particularly within the integrated SPS items. The instrument assessed SPS through written responses and therefore primarily captured students' scientific reasoning and procedural planning rather than their direct performance in laboratory settings. All responses were also scored by a single rater; although the same analytic rubric was applied consistently, inter-rater reliability could not be examined. Considering these limitations, the findings should be viewed as an initial profile of students' SPS and as a basis for further instrument refinement before broader use.

CONCLUSION

The written SPS assessment provided a preliminary diagnostic profile of pre-service science teachers' responses across 13 SPS indicators in the context of General Biology. The mean score was 22.959 out of 52 (44.151%), and several items related to data interpretation, model formulation, and hypothesis formulation showed relatively high Rasch difficulty estimates. However, because each indicator was represented by only one item, all items used the same shared stimulus on seedling growth and development, and several measurement features still require refinement, the findings should not be interpreted as a definitive representation of students' overall SPS. The results primarily reflect written SPS reasoning within the context of seedling growth and development, rather than students' direct laboratory performance. Overall, this study provides preliminary diagnostic evidence for identifying relatively challenging SPS areas and serves as a basis for refining the items, scoring categories, stimulus design, and assessment procedures before broader validation and subsequent instructional research.

RECOMMENDATION

Future studies should refine the SPS instrument by increasing the number and variety of items for each indicator, using multiple biological stimuli, clarifying the 0–4 scoring descriptors, and involving more than one rater. Written assessments should also be combined with laboratory performance measures and tested with larger, more balanced samples.

ACKNOWLEDGMENTS

The authors thank Universitas Pendidikan Mandalika for supporting the implementation of this study. Appreciation is also extended to the Biology Education, Chemistry Education, and Physics Education programs, as well as to all pre-service science teachers who willingly participated in the data collection. The authors also

thank all parties who provided academic, technical, and administrative support during instrument development, data collection, analysis, and completion of this article.

FUNDING INFORMATION
This research received no external funding. The APC was funded by the author's personal funds.

AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ahmad Aris Arifin			✓		✓	✓	✓	✓	✓		✓		✓	
Agus Ramdani	✓	✓		✓						✓		✓		
A.Wahab Jufri	✓	✓		✓						✓		✓		
Satutik Rahayu	✓	✓		✓						✓		✓		
Aris Doyan				✓						✓		✓		
A.A Sukarso	✓	✓		✓						✓		✓		

CONFLICT OF INTEREST STATEMENT
Authors state no conflict of interest.

INFORMED CONSENT
We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL
Permission to conduct the study was obtained from the participating institution. The research procedures were conducted in accordance with applicable institutional policies. Students' data were kept confidential and used solely for research purposes.

DATA AVAILABILITY
The data supporting the findings of this study are available from the corresponding author upon reasonable request. The data are not publicly available due to privacy restrictions relating to the student participants.

REFERENCES

Adlim, M., Nuzulia, R., & Nurmaliah, C. (2018). The effect of conventional laboratory practical manuals on pre-service teachers' integrated science process skills. *Journal of Turkish Science Education, 15*(4), 116–129. https://doi.org/10.12973/tused.10250a

Artayasa, I. P., Muhlis, M., Merta, I. W., Sukarso, A., & Hadiprayitno, G. (2023). Open inquiry practicum: An effective strategy for enhancing science process skills of prospective biology teachers. *Jurnal Penelitian Pendidikan IPA, 9*(3), 1352–1359. https://doi.org/10.29303/jppipa.v9i3.3248

Astalini, Darmaji, Kurniawan, D. A., Wirayuda, R. P., Putri, W. A., Rini, E. F. S., Ginting, A. A. B., & Ratnawati, T. (2023). Impact of science process skills on thinking skills in rural and urban schools. *International Journal of Instruction, 16*(2), 803–822. https://doi.org/10.29333/iji.2023.16242a

Bahtiar, B., & Dukomalamo, N. (2019). Basic science process skills of biology laboratory practice: Improving through discovery learning. *Biosfer: Jurnal Pendidikan Biologi, 12*(1), 83–93. https://doi.org/10.21009/biosferjpb.v12n1.83-93

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE—Life Sciences Education, 15*(4), rm4. https://doi.org/10.1187/cbe.16-04-0148

Hacıeminoğlu, E., Yıldız, N. G., & Şeker, R. (2022). Factors related to cognitive reasoning of pre-service teachers' science process skills: Role of experiments at home on meaningful learning. *Sustainability, 14*(13), 7703. https://doi.org/10.3390/su14137703

Handayani, S. L. (2021). Comparison of basic science process skills for students on electrical materials with the Rasch model analysis. *Jurnal Penelitian & Pengembangan Pendidikan Fisika, 7*(1), 73–82. https://doi.org/10.21009/1.07108

Harahap, F., Nasution, N. E. A., & Manurung, B. (2019). The effect of blended learning on student's learning achievement and science process skills in plant tissue culture course. *International Journal of Instruction, 12*(1), 521–538. https://doi.org/10.29333/iji.2019.12134a

Hiğde, E., & Aktamış, H. (2022). The effects of STEM activities on students' STEM career interests, motivation, science process skills, science achievements and views. *Thinking*

Skills and Creativity, 43, 101000. https://doi.org/10.1016/j.tsc.2022.101000

Hiğde, E., Yüzüak, A. V., Öcal, Z. M., & Aktamış, H. (2024). A many-facet Rasch measurement approach to analyze the prepared science laboratory activities based on science process skills and views of pre-service science teachers. *Journal of Baltic Science Education, 23*(4), 641–657. https://doi.org/10.33225/jbse/24.23.641

Idris, N., Talib, O., & Razali, F. (2022). Strategies in mastering science process skills in science experiments: A systematic literature review. *Jurnal Pendidikan IPA Indonesia, 11*(1), 155–170. https://doi.org/10.15294/jpii.v11i1.32969

Indriani, D., & Mercuriani, I. S. (2019). Experiential learning model with mind mapping on fungi: How to improve science process skills? *Biosfer: Jurnal Pendidikan Biologi, 12*(2), 223–237. https://doi.org/10.21009/biosferjpb.v12n2.223-237

Jannah, M., Oviana, W., Hayati, Z., Hidayat, R., Usman, J., & Noris, M. (2025). Enhancing pre-service teachers' science process skills through open-ended and guided inquiry-based learning. *Jurnal Ilmiah Peuradeun, 13*(2). https://doi.org/10.26811/peuradeun.v13i2.1174

Mufida, S. N., Sigit, D. V., & Ristanto, R. H. (2020). Integrated project-based e-learning with science, technology, engineering, arts, and mathematics (PjBeL-STEAM): Its effect on science process skills. *Biosfer: Jurnal Pendidikan Biologi, 13*(2), 183–200. https://doi.org/10.21009/biosferjpb.v13n2.183-200

Mustafa, N., Khairani, A. Z., & Ishak, N. A. (2021). Calibration of the science process skills among Malaysian elementary students: A Rasch model analysis. *International Journal of Evaluation and Research in Education, 10*(4), 1344–1351. https://doi.org/10.11591/ijere.v10i4.21430

Nunaki, J. H., Siagian, S. I. R., Nusantari, E., Kandowangko, N. Y., & Damopolii, I. (2020). Fostering students' process skills through inquiry-based science learning implementation. *Journal of Physics: Conference Series, 1521*, 042030. https://doi.org/10.1088/1742-6596/1521/4/042030

Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research, 15*(2), 020111. https://doi.org/10.1103/PhysRevPhysEducRes.15.020111

Rahayu, W., Putra, M. D. K., Rahmawati, Y., Hayat, B., & Koul, R. B. (2021). Validating an Indonesian version of the What Is Happening in This Class? questionnaire using a multidimensional Rasch model. *International Journal of Instruction, 14*(2), 919–934. https://doi.org/10.29333/iji.2021.14252a

Rusmini, Suyono, & Agustini, R. (2021). Analysis of science process skills of chemical education students through self-project based learning in the COVID-19 pandemic era. *Journal of Technology and Science Education, 11*(2), 371–387. https://doi.org/10.3926/jotse.1288

Suryanti, Widodo, W., & Budijastuti, W. (2020). Guided discovery problem-posing: An attempt to improve science process skills in elementary school. *International Journal of Instruction, 13*(3), 75–88. https://doi.org/10.29333/iji.2020.1336a

Susanti, R., & Anwar, Y. (2018). Profile of science process skills of pre-service biology teacher in general biology course. *Journal of Physics: Conference Series, 1006*, 012003. https://doi.org/10.1088/1742-6596/1006/1/012003

Wardani, I., & Djukri, D. (2019). Teaching science process skill using guided inquiry model with starter experiment approach: An experimental study. *JPBI (Jurnal Pendidikan Biologi Indonesia), 5*(2), 277–284. https://doi.org/10.22219/jpbi.v5i2.8429

Wola, B. R., Rungkat, J. A., & Harindah, G. M. D. (2023). Science process skills of prospective science teachers in practicum activity at the laboratory. *Jurnal Inovasi Pendidikan IPA, 9*(1), 50–61. https://doi.org/10.21831/jipi.v9i1.52974