JURNAL BIOEDUIN

p-ISSN : 2338-7173
e- ISSN : 2615-0417
(Februari 2026), Vol.(16), No.(1)

Validity of Performance Assessment of Science Process Skills in Senior High School Biology Subject

Aulia Hermawati Ulfah*[1], Supahar[2]
[1,2] Educational Research and Evaluation, Graduate School, Yogyakarta State University, Yogyakarta, Indonesia
aulia.hermawatiulfah@gmail.com*, supahar@uny.ac.id

Abstract

This study aims to develop an assessment instrument for measuring science process skills among 10th-grade high school students in Biology. The research adopts a quantitative approach and involves a sample of 415 high school students from Sumedang Regency. Participants were selected from four high schools based on the 2022 Minimum Competency Assessment (AKM) evaluation criteria. Data analysis included content validity using Aiken's V and construct validity through Confirmatory Factor Analysis (CFA). A total of 36 items were validated by 6 experts, including 2 measurement specialists, 2 subject-matter experts, and 2 high school Biology teachers. Based on the feedback, validators revised several items and removed one, resulting in 35 items being tested. The Aiken's V validity index was >0.78. Trial data were used to establish construct validity, with CFA tested against Goodness-of-Fit criteria, including GFI, RMSR/RMR, TLF/NNFI, CFI, IFI, AGFI, and RMSEA. After model modification, all items met the fit criteria for 30 items and were proven valid for assessing science process skills among 10th-grade high school students in Biology.

Keywords: Aiken'V, biology, confirmatory factor analysis, validity

INTRODUCTION

The Merdeka Curriculum, as outlined in the Ministry of Education and Culture Regulation No. 5 of 2022, stipulates that graduate competencies encompass the domains of attitudes, knowledge, and psychomotor skills. Mapping the pathways to achieve these competencies in Biology education must be supported by research and investigative skills. These competencies align with the essence of natural science, which encompasses attitudes, processes, products, and applications (Chiappetta & Koballa, 2010). Therefore, Biology education should not only focus on memorizing concepts, laws, and theories but also emphasize skills in exploration and investigation (Darmawan et al., 2021). The Merdeka Curriculum identifies two essential elements in Biology education: understanding and process skills (Ministry of Education, Culture, Research, and Technology, 2022). These process skills refer to abilities related to scientific methods or research and investigation, also known as science process skills (SPS). Training in these skills needs improvement, as research results indicate low quality.

A student can be considered proficient in science process skills when they have mastered both basic and integrated indicators (Chiappetta & Koballa, 2010; D. J. Martin, 2009; R. Martin et al., 2005; Rezba et al., 1995). Basic skills in SPS encompass the fundamental tools scientists use to gather information and understand the characteristics of the subjects they study (Khamhaengpol et al., 2021). These included observing, classifying, communicating, inferring, predicting, and measuring.

However, to achieve a deeper understanding and independently apply the scientific method, students also need to develop integrated skills that enhance basic skills (Maison et al., 2019). These skills include the ability to interpret data and design investigations.

Observing involves using senses or instruments to identify the characteristics of an object or event, while inferring refers to the ability to provide explanations based on observations (Bass et al., 2009; Chiappetta & Koballa, 2010; R. Martin et al., 2005). Predicting is the ability to forecast future events based on patterns or prior data (Chakraborty & Kidman, 2021). Meanwhile, measuring involves using appropriate tools and units to quantify an object (Chiappetta & Koballa, 2010), and classifying focuses on grouping objects or events based on their similarities and differences (Chakraborty & Kidman, 2021; Fitrianingrum & Noor, 2022). These skills culminated in communicating the information effectively, whether verbally, in writing, or through graphs and tables (Rezba et al., 1995: 15).

Integrated science process skills complement basic skills by requiring the development of analytical and planning skills. Interpreting involves the ability to explain data by analyzing patterns of relationships between variables (Foster, 1999: 128), while designing investigations entails systematic steps such as formulating objectives, establishing products, and developing tools and data analysis techniques (Rustaman, 2005; Supahar & Prasetyo, 2015). Basic and integrated science process skills are essential for preparing students to face the challenges of inquiry-based and experimental learning.

Several studies have concluded that the science process skills of high school students in Indonesia remain relatively low. Research by Agus Kurniawan et al. (2020) found that the science process skills of the 11th-grade students in Jambi were still in the "bad" category, while Mahmudah et al. (2019) reported similar findings in Bandung. Studies by Elvanisi et al. (2018) and Fitriana et al. (2019) also indicated that the science process skills of high school students need continuous improvement. Therefore, efforts to enhance these skills are essential. According to Bichi et al. (2019), Ilma et al. (2021), and Kriswantoro et al. (2021), teaching and assessment are inseparable components, and the availability of high-quality assessment instruments can support the improvement of these skills. However, in practice, various challenges arise in assessing science process skills. Several studies on the assessment of science process skills have employed performance assessment methods using direct observation techniques. Although this technique is commonly used, it has limitations in terms of space, time, and the number of assessors required. (Zuhera & Habibah, 2017) stated that time constraints, a large number of students, and the numerous indicators that need to be assessed often pose obstacles in the assessment process.

Various assessment instruments have been developed to measure science process skills, either through observation or written tests. The instrument development in this study addresses the limitations of time and space in conducting performance assessments of science process skills through observation by implementing the findings of Ruiz-Primo & Shavelson (1996). Based on findings from Ruiz-Primo & Shavelson (1996), Baxter & Shavelsons (1994), and Pine & Baxter (1991), performance assessment can be conducted through observation, notebook method, computer simulations, and paper-and-pencil tests. Therefore, the instrument in this study uses a written test to replace the performance assessment, which is usually conducted through observation. In line with this, Bass et al. (2009) state that science process skills in inquiry procedures can be measured using these methods. This type of assessment is called a confirmatory assessment because it is conducted after completing a unit of study. In other words, a written test-based performance assessment serves as a confirmatory evaluation of students' mastery of concepts and science process skills. The performance assessment instrument using this method will overcome spatial and temporal limitations, as it can be implemented widely and simultaneously.

A good instrument accurately represents students' abilities. An assessment is considered accurate when it minimizes errors or has a very low error rate (Azwar, 2022). The quality of instruments is analyzed in terms of validity, reliability, and item characteristics (Allen & Yen, 1979; Azwar, § 2022; Haryanto, 2020; Istiyono, 2020). In developing an instrument, the study must analyze at least two types of validity: content validity and construct validity. Content validity is defined as the extent to which a test evaluates the coverage of the substance intended to be measured, as assessed by expert judgment (Haryanto, 2020). On the other hand, construct validation is a procedure to prove the degree to which an instrument reflects the theoretical construct it aims to measure (Retnawati, 2016). This study reports on the validity of developing a performance assessment for science process skills among high school students in Biology.

RESEARCH METHOD

The instrument for assessing science process skills performance was developed through the ADDIE model (analyze, design, develop, implement, and evaluate) by Branch (2009), with modifications incorporating the instrument development stages from Mardapi (2016) to meet the research needs. The stages are as follows:

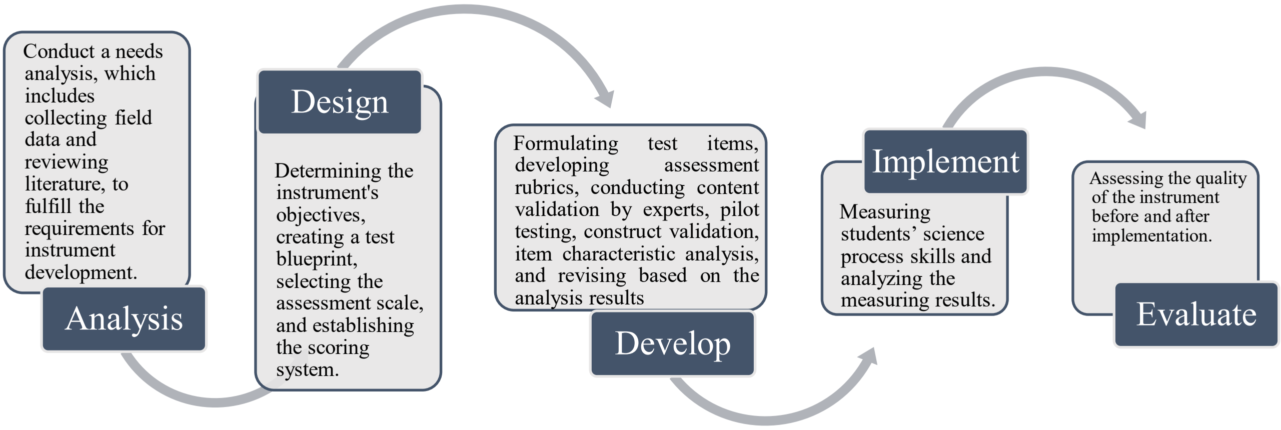

Conduct a needs analysis, which includes collecting field data and reviewing literature, to fulfill the requirements for instrument development.
Analysis
Design
Determining the instrument's objectives, creating a test blueprint, selecting the assessment scale, and establishing the scoring system.
Formulating test items, developing assessment rubrics, conducting content validation by experts, pilot testing, construct validation, item characteristic analysis, and revising based on the analysis results
Develop
Implement
Measuring students' science process skills and analyzing the measuring results.
Assessing the quality of the instrument before and after implementation.
Evaluate

Figure 1. Instrument Development Stages with the ADDIE Model

The instrument in this study was designed to assess students' knowledge retention after conducting investigations. The items were semi-divergent, using a multiple-choice format (Subali, 2009; Supahar & Prasetyo, 2015). Teachers asked students to select three out of five options that were most relevant to their post-investigation experience. Initially, the instrument consisted of 36 items developed from science process skills indicators provided by several experts (Chiappetta & Koballa, 2010; D. J. Martin, 2009; R. Martin et al., 2005; Rezba et al., 1995). The indicators in question are: observing, classifying, communicating, inferring, predicting, measuring, interpreting, and designing investigations. The maximum score a student could achieve was 4 for selecting 3 correct answers, 3 for 2 correct answers, 2 for 1 correct answer, and 1 for no correct answers. The assessment criteria were outlined in a pre-prepared rubric.

The trial involved 415 students from 4 junior high schools in Sumedang Regency. The schools were selected using a purposive sampling technique, with criteria based on the results of the 2022 Minimum Competency Assessment (AKM). The selection categories were: 1 school performing below the minimum competency level, 2 schools achieving the minimum competency level, and 1 school performing above the minimum competency level. The students completed a 35-item test using paper and pencil.

The data analysis reported includes content validity and construct validity. Contest validity involves six experts (assessment specialists, subject-matter experts, and high school Biology teachers) evaluating aspects of the material, construction, and language to ensure alignment with the development objectives. Each expert provided their assessment through the checklist and a 4-category Likert scale questionnaire (4 = item does not need revision; 3 = item requires minor revision; 2 = item requires major revision; 1 = item is unused) to avoid neutral responses. The questionnaire scores were estimated using Aiken’s V validation index, as proposed by Aiken (1985). An item was considered valid if its validation value was above 0.78. The formula for Aiken’s V validation index is as follows:

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

Meanwhile, construct validity was analyzed using Confirmatory Factor Analysis (CFA). CFA confirmed that the instrument items were converted into indicator variables derived from talent variables (Kline, 2016). CFA determines whether a construct, based on theoretical foundations, is valid. The validity of the construct/model in CFA is assessed using the Goodness-of-Fit Test (GOF). According to Hair et al. (2010), the model fit criteria should be evaluated using at least 4-5 goodness-of-fit criteria. The following criteria can represent the suitability of the CFA instrument:

Table 1. Goodness of Fit Test (GOF) Criteria

Jenis Ukuran GOF	***Cut of Value***
GFI	≥ 0,90
RMSR/RMR	≤ 0,05
RMSEA	≤ 0,08
TLI/NNFI	≥ 0,90
AGFI	≥ 0,90
CFI	≥ 0,90
IFI	≥ 0,90

In addition, item validity testing was conducted to determine the suitability of the items for the measurement model. Item validity can be assessed using the loading factor (Hair et al., 2010), with a minimum threshold of 0.5 (Yamin, 2021).

RESULT AND DISCUSSION

The development of this instrument aims to measure the science process skills of 10th-grade high school students in the Biology Subject. The test blueprint was constructed based on learning outcomes, learning objectives, and science process skills synthesized from various experts. The science process skills to be measured include basic skills: observing, inferring, conducting, classifying, predicting, measuring, and communicating; and integrated skills: interpreting and designing investigations.

After completing the instrument specification stage, constructing the test blueprint, writing the items, and developing the scoring rubric, 36 items were presented to the expert for content validation. The items were revised as necessary and subsequently piloted.

Content Validity

Content validity constitutes a crucial initial stage in the development of measurement instruments. It must be established before the instrument's implementation in field testing or subsequent statistical analysis (Roebianto et al., 2023). Six experts validated the assessment instrument developed here, and the results were analyzed using Aiken's V formula in Microsoft Excel. The Aiken's V index values are presented in Table 2.

Table 2. Results of Aiken's V Validation Calculations

Indicator	Sub-Indicator	Code	Index Aiken's V
Observing	• Determine sensory tools appropriate for the observation purpose	Item_1	0,67
	• Determine observation tools/ instruments according to their functions	Item_2 Item_3	0,61 0,89
Inferring	• Providing explanations based on the observation results of an object or event	Item_4 Item_5	0,89 0,72
	• Identifying whether a conclusion can be accepted, modified, or rejected	Item_6	0,61
Predicting	• Forecasting a pattern/ event that may occur based on previous observations and inferences	Item_7	0,89
	• Making predictions based on data patterns	Item_8	0,89
Measuring	• Selecting appropriate measuring tools for their intended use	Item_9	0,67
	• Determining measuring tools based on the measurement purpose	Item_10	0,89
	• Determining appropriate units for measuring volume, mass, or height accurately	Item_11	0,89
Classifying	• Determining the basis for grouping or categorization	Item_12	0,89
	• Identifying similarities and differences among a set of subjects/events	Item_13 Item_14	0,89 0,78
	• Grouping a set of objects/events based on observed similarities and differences	Item_15 Item_16	0,83 0,89
Communicating	• Presenting observation/investigation results orally or in writing	Item_17	0,83
	• Presenting observation/investigation results using images, graphs, diagrams, symbols, or other visual aids	Item_18	0,89
Interpreting	• Analyzing patterns and relationships among observed/investigated facts	Item_19	0,89

Indicator	Sub-Indicator	Code	Index Aiken's V
	• Providing explanations based on the analysis of two or more inferences	Item_20	0,89
	• Relating observation results to existing theories/concepts	Item_21	0,89
		Item_22	0,83
Designing Investigation	• Formulating the purpose of the investigation	Item_23	0,89
	• Formulating the benefits of the investigation	Item_24	0,89
	• Determining the tools/materials/resources used according to the investigation purpose	Item_25	0,89
		Item_26	0,94
	• Describing the investigation procedure	Item_27	0,94
		Item_28	0,89
	• Determining what will be measured, observed, and recorded	Item_29	1,00
		Item_30	0,94
	• Establishing data collection procedures for the investigation	Item_31	0,94
		Item_32	1,00
	• Designing the setup of the investigation equipment	Item_33	0,94
	• Designing the presentation of investigation results	Item_34	1,00
		Item_35	0,89
	• Designing data analysis techniques for investigating results	Item_36	1,00

Based on Table 2, several items were deemed invalid as they had an Aiken's V index value below 0.78. these items include item_1, item_2, item_5, item_6 and item_9. Invalid items can be revised according to the validators' suggestions. Necessary improvements may include refining the wording of the questions, adjusting the distractions, and using more precise language aligned with the measurement objective. The narratives for Item_1 and Item_2 were revised because they remained ambiguous and contained repetitive words within a single sentence, thereby reducing the readability of the questions. However, communicative language is a crucial principle in question design (Pusat Penilaian Pendidikan, 2016). Communicative sentences help students more easily grasp the information or ideas presented in the questions (Rahmawati et al., 2021). Effective sentence construction, supported by appropriate diction and correct punctuation, enhances communicative clarity (Akbar, 2020). With these improvements, students will better understand the intent of the questions. Ambiguous sentence structures or editorial complexities that are not directly related to the assessed construct may obscure the assessment's purpose and place greater emphasis on students' linguistic proficiency or guessing strategies rather than on their conceptual understanding (Liu et al., 2024).

On the other hand, the answer choices for item 2 were less effective as distractors, making it likely that students could easily guess the incorrect answers even if their abilities were low. Developing well-functioning distractors is a challenge in the construction of multiple-choice instruments; however, to obtain high-quality items, the answer options must include distractors that are logically homogeneous in terms of content and grammatical structure (Stevens et al., 2023). Distractions are designed to mislead students, particularly those with a limited understanding of the material, into selecting them (Qadir et al., 2024). If an option is distinguishable from the others, it becomes easy to identify as either correct or incorrect. Therefore, distractors must be carefully crafted to resemble the correct answer but not be identical to it, thereby increasing the likelihood of being chosen (Stevens et al., 2023). Efficient distractors can influence item difficulty and discrimination indices; therefore, improving the quality of distractors may serve as a strategic approach to developing valid and reliable assessments (Oc & Hassen, 2025)

One of the validators suggested revising Item 6 because the question's narrative did not align with the measured indicator, which is inferencing. Inferencing is the skill of providing explanations based on observations of an object or event (Chiappetta & Koballa, 2010). In contrast, Item 6 asked students to assess data validity, which does not align with the definition of inferencing. Therefore, Item 6 was revised to better align with the indicator. The question's narrative or wording should guide students to demonstrate their abilities in the measured variable, as this is evidence of a valid item (Allen & Yen, 1979). If an item directs students toward a different skill, the test developer must redefine the measured variable. Constructing items that are misaligned with the intended indicators or measurement levels risks obscuring the assessment's purpose, resulting in shallow, non-representative measurement (Liu et al., 2024).

As for Item 9, changes were made only to the wording of the answer choice. All answer choices for Item 9 began with the word "*mengukur*" (to measure). The validator noted that this was neither effective nor efficient, and suggested moving the word "*mengukur*" to the question stem. This would prevent students from repeatedly encountering the word while reading the answer choices. This aligns with the principle of question writing, which states that repetition of the same word or phrase in all options should be avoided unless it serves a specific purpose or function in the context of the question (Pusat Penilaian Pendidikan, 2016).

Additionally, the validator recommended revising the sub-indicator for the "measuring" indicator from "using measuring tools according to their functions" to "Selecting appropriate measuring tools for their intended use." Consequently, the indicator for Item 9 was revised, and the question's narrative was adjusted to more specifically measure the intended sub-indicator. The question stem must align with the indicator, and the indicator must align with the material being assessed (Pusat Penilaian Pendidikan, 2016).

On the other hand, Item 5 was proposed for removal because, according to Biology teachers, its content was less suitable for high school students. The removal of item 5 did not disrupt the instrument's construction, as there were still other items representing the same indicator. Other invalid items were revised based on the validators' suggestions. Meanwhile, valid items aligned with the intended indicators, content, and context. Thus, 35 items were used in the trial

Construct Validity

The instrument was developed based on a thorough theoretical review, so construct validity in this study was demonstrated using Confirmatory Factor Analysis (CFA). The first step was to examine the fit between the instrument model and empirical field data using the Goodness of Fit Test (GOF) criteria. Table 3 presents the GOF values for the developed instrument, estimated using the Lisrel 8.5 application.

Table 3. Result of the GOF Test for 35-Item Science Process Skills Performance Assessment Instrument

Types GOF	Cut of Value (*Nilai Batas*)	Result of GOF Value	Inference
GFI	≥ 0,90	0,88	Marginal fit
RMSR/RMR	≤ 0,05	0,073	Miss fit
RMSEA	≤ 0,08	0,043	Close fit
TLI/NNFI	≥ 0,90	0,94	Good Fit
AGFI	≥ 0,90	0,86	Marginal fit
CFI	≥ 0,90	0,95	Good Fit
IFI	≥ 0,90	0,95	Good Fit

The criteria for GFI, TLI/NNFI, AGFI, CFI, and IFI are considered a good fit if their values are ≥ 0.90, and a marginal fit is within the range of 0.80 to 0.90. The criteria for RMSR/RMR are considered a good fit if the value is ≤ 0.05, while the criterion for RMSEA is considered a good fit if the value is ≤ 0.08 and a close fit if the value is < 0.05 (Haryono, 2016). Based on these criteria, the 35-item assessment developed in this study is deemed fit for the TLF/NNFI, CFI, and IFI criteria; marginal fit for the GFI and AGFI criteria; close fit for the RMSEA criterion; and not fit for the RMSR/RMR criterion.

Furthermore, the suitability of each item to the instrument's construct is assessed using the loading factor. If the loading factor of an item is ≥ 0.5, the item is considered valid, and vice versa (Haryono, 2016). The loading factor values for each item to the dimensions, and from the dimensions to the variable, are presented in Figure 1.

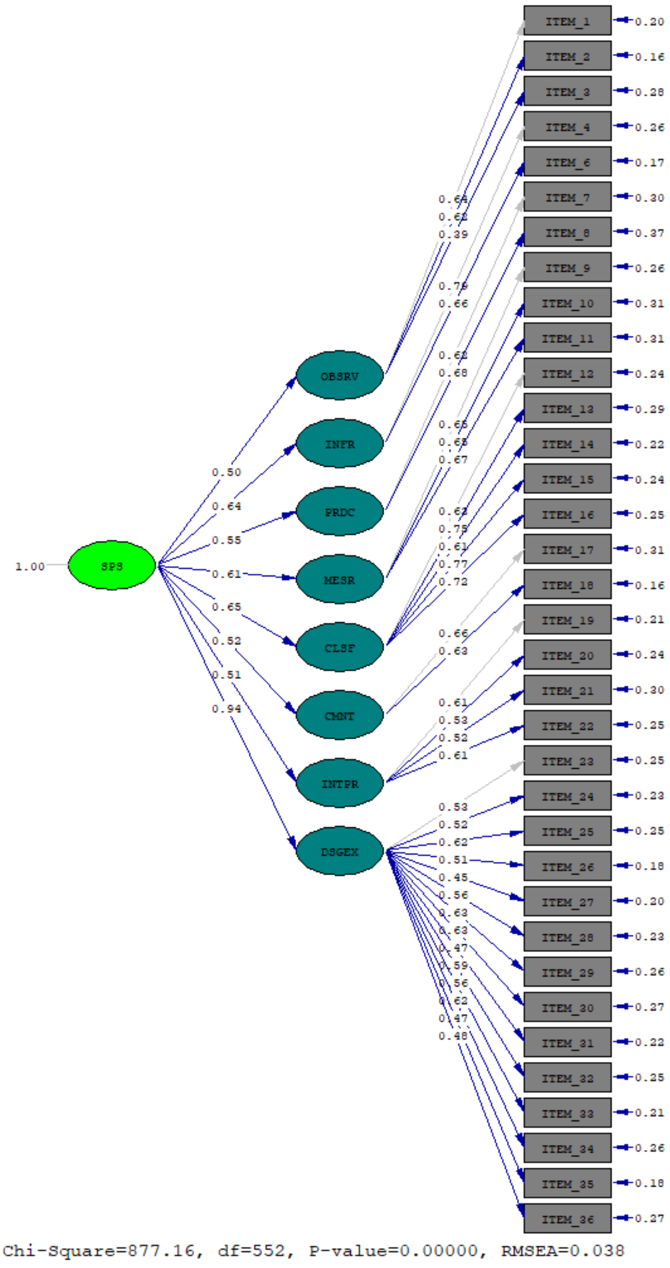
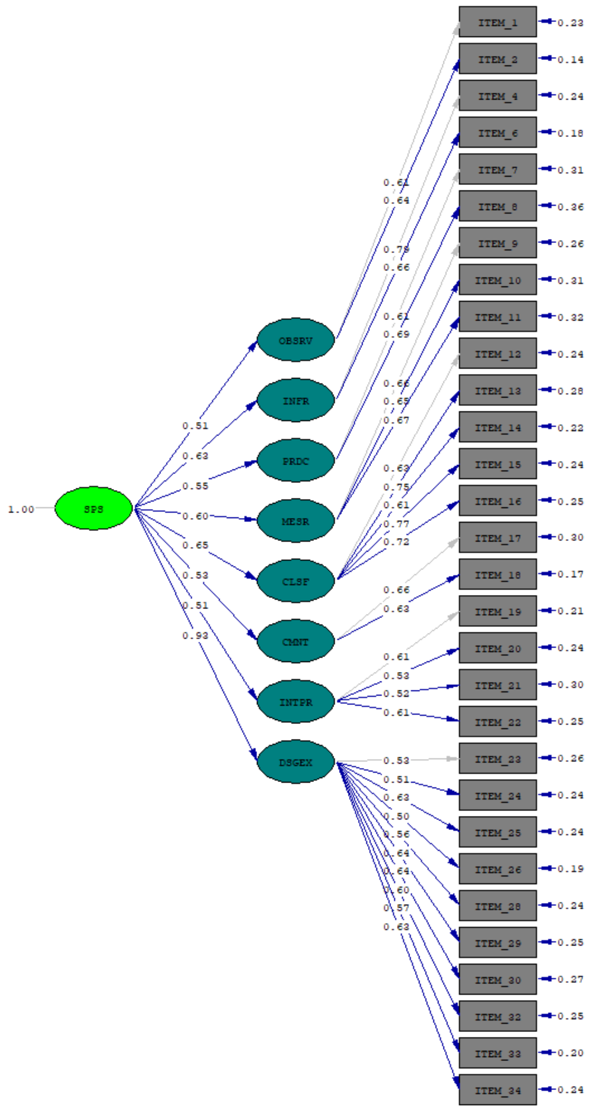


Figure 2. Path Diagram of CFA for the 35-Item Science Process Skills Performance Assessment Instrument

Based on Figure 1, there are five items with a loading factor (LF) value < 0.5, namely item_3, item_27, item_31, item_35, and item_36. These items did not meet the required loading factor threshold and can therefore be declared invalid for the model. This issue may be influenced by unclear question wording, irrelevant content, or poorly constructed answer choices. We modified the model by eliminating items with loading factor values below 0.5. According to Rahmania et al. (2022), measurement model modifications can be made to achieve a better model. The results of the CFA analysis for the modified 30-item model are presented in Table 4 and Figure 2.

Table 4. Result of the GOF Test for the 30-Item Science Process Skills Performance Assessment Instrument

Types GOF	***Cut of Value*** **(Nilai Batas)**	Result of GOF Value	Inference
GFI	≥ 0,90	0,91	Good Fit
RMSR/RMR	≤ 0,05	0,027	Good Fit
RMSEA	≤ 0,08	0,037	Close Fit
TLI/NNFI	≥ 0,90	0,96	Good Fit
AGFI	≥ 0,90	0,89	Marginal fit
CFI	≥ 0,90	0,97	Good Fit
IFI	≥ 0,90	0,97	Good Fit



Chi-Square=623.42, df=397, P-value=0.00000, RMSEA=0.037

Figure 3. Path Diagram of CFA for the 30-Item Science Process Skills Performance Assessment Instrument

Items with a loading factor value of less than 0.5 are considered unsuitable for measuring students' science process skills. After modifications, the new model demonstrated improved GOF. The model achieved a good fit for the GFI, RMSR/RMR, TLF/NNFI, CFI, and IFI criteria, a marginal fit for the AGFI criterion, and a close fit for the RMSEA criterion. According to Haryono (2016), a model is considered acceptable if it meets at least one fit criterion; the greater the number of fit criteria, the better the model. The science process skills performance assessment instrument met several fit criteria, confirming its validity. Additionally, the new model was re-estimated using Lisrel 8.5, with all items achieving loadings greater than 0.5. Based on these results, it can be concluded that the science process skills performance assessment model is acceptable (fit) with a total of 30 validated items, namely: item_1, item_2, item_ 4, item_6, item_7, item_8, item_9, item_10, item_11, item_12, item_13, item_14, item_15, item_16, item_17, item_18, item_19, item_20, item_21, item_22, item_23, item_24, item_25, item_26, item_28, item_29, item_30, item_32, item_33, dan item_34.

This instrument can serve as a reference or model for implementing other assessments, particularly in Biology learning. Teachers who face challenges in conducting authentic assessments in the classroom, such as limited time and large class sizes, can develop similar instruments as an alternative assessment method. The ease of administering performance assessments using the paper-and-pencil test method is expected to encourage teachers to be more creative in designing more efficient assessments.

CONCLUSION

The science process skills performance assessment instrument in this study initially consisted of 36 items. Based on content validation, four items required revision, including clarifying ambiguous sentences, replacing ineffective distractors, modifying questions that were misaligned with the indicators, and eliminating one irrelevant item. Consequently, 35 items were confirmed valid through content validation. The instrument was further tested using CFA. The instrument met the CFA testing standards based on the Goodness-of-Fit Criteria, including GFI, RMSR/RMR, TLF/NNFI, CFI, IFI, AGFI, and RMSEA. However, only 30 items met the required loading factor threshold. Therefore, the final validated instrument consists of 30 items, capable of measuring the science process skills of tenth-grade high school students in the Biology subject.

BIBLIOGRAPHY

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

Akbar, A. (2020). Kemampuan mahasiswa dalam penyusunan soal pilihan ganda. *Attadib Journal Of Elementary Education*, *4*(1), 44–53. https://www.jurnalfai-uikabogor.org/index.php/attadib/issue/view/52

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

Azwar, S. (2022). *Reliabilitas dan validitas*. Pustaka Belajar.

Bass, J. E., Contant, T. L., & Carin, A. A. (2009). *Teaching scienceas inquiry* (8th ed.). Allyn & Bacon.

Bichi, A. A., Ibrahim, R. H., & Ibrahim, F. B. (2019). Assessment of students performances in biology: Implication for measurements and evaluation of learning. *Journal of Education and Learning (EduLearn)*, *13*(3), 301–308. https://doi.org/10.11591/edulearn.v13i3.12200

Branch, R. M. (2009). *Instructional Design: The ADDIE Approach*. Springer Science & Business Media.

Chakraborty, D., & Kidman, G. (2021). Inquiry process skills in primary science textbooks: authors and publishers' Intentions. *Research in Science Education*. https://doi.org/10.1007/s11165-021-09996-4

Chiappetta, E. L., & Koballa, T. R. (2010). *Science instruction in the middle and secondary school* (7th ed.). Pearson Education.

Darmawan, E., Yusnaeni, Ismirawati, N., & Riswanto, R. H. (2021). *Strategi belajar mengajar Biologi*. Pustaka Rumah C1nta.

Elvanisi, A., Hidayat, S., & Fadillah, E. N. (2018). Analisis keterampilan proses sains siswa sekolah menengah atas. *Jurnal Inovasi Pendidikan IPA*, *4*(2), 245–252. https://doi.org/10.21831/jipi.v4i2.21426

Fitriana, F., Kurniawati, Y., & Utami, L. (2019). Analisis keterampilan proses sains peserta didik pada materi laju reaksi melalui model pembelajaran bounded inquiry laboratory. *JTK (Jurnal Tadris Kimiya)*, *4*(2), 226–236. https://doi.org/10.15575/jtk.v4i2.5669

Fitrianingrum, H., & Noor, M. F. (2022). Science process skills students on cells and tissues concept during the covid-19 pandemic: How did it achieve? *Journal of Physics: Conference Series*, *2157*(1). https://doi.org/10.1088/1742-6596/2157/1/012041

Hair, J. F., Black, W. C., Babin, B. J., & Anserson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson Education.

Haryanto. (2020). *Evaluasi pembelajaran (konsep dan manajemen)*. UNY Press.

Ilma, A. Z., Adhelacahya, K., & Ekawati, E. Y. (2021). Assessment for learning model in competency assessment of 21st century student assisted by google classroom. *Journal of Physics: Conference Series*, *1805*(1). https://doi.org/10.1088/1742-6596/1805/1/012005

Istiyono, E. (2020). *Pengembangan instrumen penilaian dan analisis hasil belajar Fisika dengan teori tes klasik dan modern* (2nd ed.). UNY Press.

Khamhaengpol, A., Sriprom, M., & Chuamchaitrakool, P. (2021). Development of STEAM activity on nanotechnology to determine basic science process skills and engineering design process for high school students. *Thinking Skills and Creativity*, *39*. https://doi.org/10.1016/j.tsc.2021.100796

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. The Guilford Press.

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

Liu, Q., Wald, N., Daskon, C., & Harland, T. (2024). Multiple-choice questions (MCQs) for higher-order cognition: Perspectives of university teachers. *Innovations in Education and Teaching International*, *61*(4), 802–814. https://doi.org/10.1080/14703297.2023.2222715

Maison, M., Darmaji, D., Kurniawan, D. A., Astalini, A., Dewi, U. P., & Kartina, L. (2019). Analysis of science process skills in physics education students. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *23*(2), 197–205. https://doi.org/10.21831/pep.v23i2.28123

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.

Martin, R., Sexton, C., & Franklin, T. (2005). *Teaching science for all children: inquiry methods for constructing understanding*.

Oc, Y., & Hassen, H. (2025). Comparing the effectiveness of multiple-answer and single-answer multiple-choice questions in assessing student learning. *Marketing Education Review*, *35*(1), 44–57. https://doi.org/10.1080/10528008.2024.2417106

Pusat Penilaian Pendidikan. (2016). *Panduan penulisan*. Kementrian Pendidikan dan Kebudayaan.

Qadir, A., Huda, N., & Hermina, D. (2024). Analisis butir tes: tingkat kesukaran, daya pembeda dan efektivitas pengecoh. *Al Furqan: Jurnal Agama, Sosial, Dan Budaya*, *3*(3), 1450–1467. https://publisherqu.com/index.php/Al-Furqan/article/view/1069/962

Rahmania, A., Triwahyuni, A., & Kadiyono, A. L. (2022). Validitas konstruk growth mindset scale: versi Bahasa Indonesia. *Jurnal Psikologi*, *18*(2), 194. https://doi.org/10.24014/jp.v18i2.16925

Rahmawati, I., Suryana, Y., & Hidayat, S. (2021). Analisis kesesuaian soal penilaian tengah semester IPA dengan kaidah penyusunan soal pada aspek bahasa di sekolah dasar. *Edukatif: Jurnal Ilmu Pendidikan*, *3*(6), 3636–3646. https://doi.org/10.31004/edukatif.v3i6.975

Retnawati, H. (2016). *Analisis kuantitatif instrumen penelitian (panduan peneliti, mahasiswa, dan psikometri)*. Parama Publishing.

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.

Roebianto, A., Savitri, S. I., Aulia, I., Suciyana, A., & Mubarokah, L. (2023). Content validity: definition and procedure of content validation in psychological research. *TPM - Testing, Psychometrics, Methodology in Applied Psychology*, *30*(1), 5–18. https://doi.org/10.4473/TPM30.1.1

Rustaman, N. (2005). *Strategi belajar mengajar Biologi*. UM Press.

Stevens, S. P., Palocsay, S. W., & Novoa, L. J. (2023). Practical Guidance for writing multiple-choice test questions in introductory analytics courses. *INFORMS Transactions on Education*, *24*(1), 51–69. https://doi.org/10.1287/ited.2022.0274

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

Yamin, S. (2021). *SmartPLS 3, AMOS, & STATA Olah data statistik (mudah & praktis)*. Dewangga Energi Internasional.

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.