JAPRA
Jurnal Pendidikan Raudhatul Athfal
P-ISSN. 2527-4325 E-ISSN. 2580-7412

Validity and Reliability of Formative Assessment in STEAM Learning Based on Mount Merapi Eruption Experiment at the Foundational Phase

Hasna Alhazimatu Shufiya[1], Nuraini[1], Tuti Hayati[2]
[1, 2,] Universitas Islam Negeri Sunan Gunung Djati
Jl. Cimencrang, Kec. Gedebage, Kota Bandung, Jawa Barat 40292
Email: hasnaalhazimatushufiya6@gmail.com[1] , aininurainii@gmail.com[1], thayati18@gmail.com[2]

Naskah diterima: Oktober 8, 2025., direvisi: Oktober 27, 2025., diterbitkan: Desember 30, 2025.

Abstrak

Instrumen penilaian formatif yang terstandar dalam pendidikan sains untuk anak usia dini masih jarang ditemukan, sehingga dapat menyebabkan bias dalam pengamatan. Penelitian ini bertujuan untuk membuktikn validitas dan reliabilitas asesmen formatif eksperimen STEAM: letusan gunung merapi fase fondasi. Dengan menggunakan Pengembangan (Research and Development/R&D) dengan fokus pada validitas dan reliabilitas instrumen. Penelitian ini melibatkan anak-anak berusia 5–6 tahun dari Kelompok B PAUD Lembut Kadudampit Sukabumi yang dipilih secara total sampling. Data dikumpulkan melalui lembar observasi terstruktur berbasis kinerja dan dianalisis menggunakan rumus Pearson Product-Moment serta Cronbach's Alpha. Hasil penelitian menunjukkan bahwa 6 item pernyataan dinyatakan valid secara empiris, dan instrumen secara keseluruhan memiliki keandalan/reliabilitas yang tinggi. Item observasi yang bias dipengaruhi oleh adanya pendampingan dalam praktik. Selain itu, pengaruh peniruan respons antar anak pada kegiatan klasikal. Beberapa item dieliminasi, sehingga instrumen penilaian ini menjadi ringkas dan objektif.

***Kata Kunci**: Asesmen, Eksperimen STEAM, Fase Fondasi, Sains*

Abstract

Standardized formative assessment instruments for early childhood science education remain scarce, a gap that can lead to observational bias. This study aimed to establish the validity and reliability of a formative assessment instrument designed for a STEAM experiment specifically, a volcanic eruption activity at the foundational education level. A Research and Development (R&D) approach was employed, focusing on the instrument's validity and reliability. Participants included children aged 5–6 years from Group B of the

Lembut Early Childhood Education (PAUD) center in Kadudampit, Sukabumi, selected via total sampling. Data were collected using structured, performance-based observation sheets and analyzed using the Pearson Product-Moment formula and Cronbach's Alpha. The results demonstrated that six statement items were empirically valid and that the instrument as a whole exhibited high reliability. Instances of biased observation items were attributed to the presence of adult guidance during the activity and the tendency for children to mimic each other's responses during group sessions. Following the elimination of certain items, the resulting assessment instrument is now concise and objective.
Keywords: Assessment, STEAM Experiments, Foundation Phase, Science

Introduction

Early Childhood Education (PAUD) at the Foundational Phase serves as a vital pillar in constructing children's intelligence, social competence, and personality. As an early form of stimulation, early childhood education integrates the fulfillment of physical and psychological growth and development needs from birth through six years of age, establishing a solid foundation that prepares children to transition into subsequent learning stages. Among various developmental learning outcomes, the introduction of basic science serves as an essential instrument for stimulating holistic growth and development. At an early age, science education is not oriented toward mastering complex theoretical concepts, but rather toward authentic processes in which children observe, explore various objects, and recognize causality in surrounding natural phenomena. Through an appropriate approach, science can spark logical reasoning, nurture natural curiosity, and construct simple problem-solving skills from an early age (Anggun et al., 2024; Sriyono, 2021).

Basic science learning at the Foundational Phase becomes far more meaningful when conducted through hands-on experiences using real objects (Nisa, 2024; Toyyib, 2024). One of the most ideal methods to achieve this is the experimental concept. Among various options, the "Mount Merapi Eruption" experiment serves as a highly effective medium because it presents real-world chemical changes while remaining safe for young children. This activity utilizes a simple acid-base reaction combining vinegar, baking soda, and liquid soap to generate carbon dioxide (CO_2) gas pressure, driving a foam overflow that resembles lava (Chintya & Andriani, 2025; Fadhila et al., 2025).

However, the intense visual appeal of this lava volcano experiment creates a new

operational need in the field, particularly regarding the standardization of children's science process assessment. Based on preliminary observations at PAUD Lembut Kadudampit Sukabumi, a recurring issue during this engaging experimental activity is how to objectively measure children's cognitive achievements. Formative assessment methods currently utilized in schools are sometimes overly general, leaving them susceptible to ceiling effect bias (Fadhila et al., 2025). A condition wherein children's enthusiasm and excitement upon witnessing the overflowing foam potentially obscure the accurate assessment of their actual cause-and-effect reasoning or object classification abilities (Chintya & Andriani, 2025).

Therefore, the development of a valid and reliable assessment instrument is essential for systematically observing children's abilities during the Foundational Phase. Assessing real-activity-based STEAM learning requires a formative assessment instrument equipped with indicators that are relevant, representative, specific, and directly observable (Fadhila et al., 2025). In the context of the "Mount Merapi Eruption" experiment, five STEAM elements Science, Technology, Engineering, Arts, and Mathematics were selected as the framework for constructing indicators due to their direct connection to the sequence of activities performed by the children. The Science element encompasses observing phenomena and occurring changes; Technology pertains to the utilization of tools and materials; Engineering involves designing the volcano model and executing the experiment; Arts relates to creativity in designing or decorating the model; and Mathematics covers the recognition of quantities, sizes, and simple comparisons.

These five elements serve as a reference point for aligning activity objectives with children's observable behaviors during the experiment, thereby allowing assessment indicators to be constructed in accordance with the learning context. Previous research by Eliamah et al. (2022) examined learning through experimental methods in early childhood education. However, testing the quality of an assessment instrument within the context of a STEAM "Mount Merapi Eruption" experiment remains necessary to obtain empirical evidence regarding the validity and reliability of its items. This instrument is expected to support teachers and researchers in conducting more systematic observations of children's science process skills, aligned with the importance of developing assessment tools congruent with learning objectives (Nugraha & Rachmawati, 2020).

The objective of this study is to examine the empirical validity and reliability of a formative assessment instrument for STEAM learning via the "Mount Merapi Eruption"

experiment among early childhood learners at the Foundational Phase. The evaluation findings are expected to provide initial evidence regarding item quality and serve as a basis for consideration in utilizing and developing instruments in future research.

Methodology

This assessment instrument was developed using a Research and Development (R&D) approach incorporating Branch's (2009) ADDIE model (analyze, design, develop, implement, and evaluate), integrated with the research instrument development stages formulated by Retnawati (2016). However, this study focuses specifically on analyzing the validity and reliability of the instrument by applying a quantitative descriptive research method. The application of this method is fully focused on testing the validity and reliability feasibility of the formative assessment instrument for the STEAM experiment: Mount Merapi eruption at the foundational phase using real field-trial data.

The "simple lava volcano" science experiment activity was utilized as a stimulus to naturally elicit children's science process skills. This research was conducted at PAUD Lembut Kadudampit Sukabumi. The population of this study comprised all children in Group B. Sampling was conducted using a total sampling technique due to the relatively small population size in Group B, resulting in all 12 children in Group B being included as the complete sample. The children's developmental characteristics corresponded to the foundational phase, with an age range of 5–6 years.

The primary data collection tool tested was a quantitative observation sheet with four developmental rating options: Not Yet Developed (BB / score 1), Starting to Develop (MB / score 2), Developing as Expected (BSH / score 3), and Very Well Developed (BSB / score 4). These primary data were reinforced with classroom anecdotal records to capture unique occurrences or specific child responses during the experiment.

The structure of the instrument was derived directly from the STEAM Learning Outcome (CP) Elements of the Independent Curriculum (Kurikulum Merdeka). These targets were detailed into 5 main indicator groups distributed across 10 statement items, including the ability to observe natural phenomena in detail to measure children's visual focus in following the sequential emergence of foam (Items 1 and 2), and the ability to predict outcomes to evaluate early logical capabilities in guessing the reaction before experiment completion (Items 3 and 4).

Furthermore, the ability to identify cause-and-effect relationships was assessed to detect children's understanding of how mixing materials causes foam overflow (Items 5 and 6). Meanwhile, classification and pattern recognition abilities were evaluated to observe children's skill in distinguishing types of materials based on their physical state and their ability to sequence experimental steps (Items 7 and 8). Finally, the ability to analyze object changes measured children's sensitivity in recognizing changes in the physical size and shape of materials (Items 9 and 10). These descriptions are presented in Table 1 below.

Table 1. Assessment Indicators

No.	Main Indicators (STEAM)	Measurement Focus / Behavioral Indicators	Item Numbers
1	Ability to Observe Natural Phenomena	Children's visual focus in following the sequential emergence of foam	1, 2
2	Ability to Predict Outcomes	Children's early logical reasoning in predicting reactions before experiment completion	3, 4
3	Ability to Identify Cause-and-Effect Relationships	Children's understanding of material mixing causing foam overflow	5, 6
4	Classification and Pattern Recognition Ability	Children's skill in distinguishing materials by physical state and sequencing experimental steps	7, 8
5	Ability to Analyze Object Changes	Children's sensitivity in identifying changes in physical size and shape of materials	9, 10

The data analysis process was conducted in two stages: validity testing and reliability testing. Item validity was tested using Pearson's Product-Moment correlation at a 5% significance level. With a sample size of 12 children, the critical value of r_table was set at 0.576; thus, an item was declared valid if r_calculated > 0.576. Subsequently, reliability testing was conducted using Cronbach's Alpha coefficient via Microsoft Excel. A reliability coefficient of 0.70 or higher served as the general benchmark for acceptable reliability, while accounting for instrument characteristics and usage context (Taber, 2018).

Results and Discussion

A field trial of the 10 formative STEAM experiment assessment items was conducted through observations of 12 Group B children at PAUD Lembut Kadudampit Sukabumi. Raw observation scores ranged from 1 to 4, based on the early childhood developmental rating scale: 1 for Not Yet Developed (BB), 2 for Starting to Develop (MB),

3 for Developing as Expected (BSH), and 4 for Very Well Developed (BSB). Individual score acquisitions are presented comprehensively in Table 2.

Tabel 2. Observation Results of Formative Assessment in Early Childhood Experiment Activities

Student	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	Total
Child 1	4	3	2	3	1	3	3	3	3	4	30
Child 2	4	3	3	3	2	3	4	2	3	1	28
Child 3	4	4	4	3	1	2	4	2	1	1	26
Child 4	4	4	4	4	4	3	4	4	3	4	38
Child 5	4	4	4	3	2	3	4	3	2	2	31
Child 6	3	3	4	4	2	3	4	3	3	2	31
Child 7	4	4	4	4	3	3	4	4	3	2	35
Child 8	4	4	4	4	3	4	4	4	3	2	36
Child 9	4	4	4	4	3	2	4	3	3	2	33
Child 10	4	1	2	2	2	4	4	4	3	1	26
Child 11	4	4	4	4	3	4	4	4	3	3	37
Child 12	4	1	2	2	3	2	4	3	3	1	25
Variance (S^2)	0.083	1.295	0.811	0.932	0.811	0.545	0.083	0.447	0.424	1.174	(S^2) total = 20.424

Based on Table 2, total student scores ranged from 25 to 38 out of a maximum score of 40. Child 4 achieved the highest score of 38 (95%), whereas Child 12 obtained the lowest score of 25 (62.5%). Most children scored above 30, indicating that the majority demonstrated excellent cognitive science process abilities during the experimental simulation.

Instrument Validity Test Results

Empirical validity analysis of the 10 observation statement items was conducted using Karl Pearson's Product-Moment correlation formula. The mathematical formula applied to determine the correlation coefficient is as follows:

$$r_{xy} = \frac{N(\sum XY) - (\sum X)(\sum Y)}{\sqrt{\{N\sum X^2 - (\sum X)^2\}\{N\sum Y - (\sum Y)^2\}}}$$

The calculated results obtained from the above formula were subsequently compared against the critical threshold value of r_table = 0.576 at a significance level of $\alpha = 0.05$ with a sample size of N = 12. The item validity testing results are presented in Table 3 below.

Table 3. Item Validity Test Results of the STEAM Assessment Tool

Item Number	Correlation Value (r_calculated)	Benchmark Value (r_table)	Item Feasibility Status

Item 1	0.023221	0.576	Invalid
Item 2	0.706940	0.576	Valid
Item 3	0.655376	0.576	Valid
Item 4	0.812705	0.576	Valid
Item 5	0.655376	0.576	Valid
Item 6	0.408551	0.576	Invalid
Item 7	0.581703	0.576	Valid
Item 8	0.092910	0.576	Invalid
Item 9	0.442664	0.576	Invalid
Item 10	0.680654	0.576	Valid

Based on Table 3, among the 10 tested instrument items, 6 items were declared valid (Items 2, 3, 4, 5, 7, and 10) as their calculated values satisfied r_calculated > 0.576. Conversely, 4 items were declared invalid—namely Item 1 (r = 0.023), Item 6 (r = 0.409), Item 8 (r = 0.093), and Item 9 (r = 0.443)—because their calculated correlation values fell below 0.576.

Instrument Reliability Test Results

Internal consistency testing for this assessment was conducted comprehensively using Cronbach's Alpha coefficient. The mathematical Cronbach's Alpha formula applied is as follows:

$$r_{11} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum s^2}{S^2_{total}}\right)$$

Data calculation based on the total item variance $\sum s_i^2 = 6.606060$ and total score variance $s_{total}^2 = 20.42424$ formulated into the equation yielded the following reliability coefficient:

$$r_{11} = \left(\frac{10}{10-1}\right)\left(1 - \frac{6{,}606060}{20{,}42424}\right) = 0{,}751730$$

The final reliability coefficient derived from this formula calculation is presented in Table 4 below.

Table 4. Reliability Estimation Results

Parameter Description	Formula Coefficient Value (r_11)	Sample Size (N)	Reliability Interpretation
Total 10 Statement Items	0.751730	12	High

Based on Table 4, the obtained Cronbach's Alpha coefficient was 0.751730. This value falls into the High reliability classification ($\alpha \geq 0.70$), demonstrating that the instrument tool is reliable, stable, and consistent for measuring early childhood scientific cognitive abilities at the Foundational Phase.

The evaluation of validity and reliability in this STEAM-based science formative assessment provides crucial insights into instrument characteristics. Through quantitative analysis of trial data, instrument properties can be objectively mapped to determine the degree to which constructed statement items capture early childhood scientific thinking capabilities without introducing observational bias.

The execution of early childhood assessment in field settings is expected to produce objective data. However, implementation challenges persist; findings by Subiani et al. (2022) highlight that field teachers still experience technical knowledge limitations in independently constructing standardized assessment instruments. Therefore, the presence of standardized instruments is critical to prevent teachers from falling back on subjective developmental evaluations.

The invalidity identified in several statement items reflects aspects requiring further refinement. For Item 1 ($r = 0.023$), this low figure is presumed to stem from score homogeneity, wherein children's score distributions clustered at the maximum value. This occurred because the statement stimulus was overly general, the measured dimension lacked specificity, or it assessed a skill uniformly attainable by children of that age. This condition demonstrates that the item lacked discriminatory power.

These results reinforce the view of Anggreni et al. (2022) that early childhood science instruments must be designed in accordance with children's developmental levels and stimulation needs. Without adequate discriminatory components, assessment tools risk generating uniform (homogeneous) scores, failing to accurately reflect variations in individual ability.

For Item 8 ("following sequential steps"), the correlation value was likewise invalid at $r = 0.093$. The low correlation was influenced by the situational and collective nature of early childhood science learning. When children perform experiments together, their tendency to imitate peers' actions prevents pure measurement of individual independence. Other factors, such as teacher guidance, also contributed to assessment bias.

Findings regarding assessment obstacles align with research by Akbar et al. (2024). In practice, intensive teacher involvement acts as a potential confounding variable, making it difficult to separate children's independent achievements from the effects of received guidance.

On the other hand, the successful testing of valid items confirms that the instrument design effectively captures science process skills. Early-stage scientific thinking abilities were captured in Item 4 (predicting overflow limits), which yielded the highest correlation value of r = 0.813. The characteristic of this item prompts children to integrate cognitive experiences to hypothesize new phenomena.

The success of this prediction item aligns with arguments in prior research on science literacy development in Group B children discussed by Westhisi et al. (2019) and Handayani & Dewi (2023). They stated that science literacy-based stimulation is proven effective in triggering critical thinking foundational skills, such as observing, measuring, predicting, and communicating natural phenomena logically.

Integrating cognitive science aspects through the "Mount Merapi Eruption" experiment affords children opportunities to develop science process skills through direct experience. This is consistent with research by Syahrul et al. (2023), demonstrating that experimental methods enhance science process skills in children aged 5–6 years, particularly in observing, predicting, and communicating experimental outcomes.

The positive impact of direct stimulation is further supported by Sari (2020), who demonstrated a significant effect of experimental methods on cognitive acuity and science ability among children aged 5–6 years (Group B). From a measurement instrument development perspective, achieving a Cronbach's Alpha coefficient of 0.752 (High category) assures that this assessment is reliable for deployment. Field-observation item characteristic testing follows standard procedures established in studies by Bashooir & Supahar (2018) and Qonita et al. (2025).

Instrument development underscores that establishing empirical validity and estimating reliability are essential prerequisites that must not be omitted, ensuring a performance assessment instrument possesses valid theoretical functionality prior to widespread application in early childhood education institutions.

Ultimately, refining the initial 10 items into 6 valid items (Items 2, 3, 4, 5, 7, and 10) adequately addresses the need for a well-characterized early childhood science evaluation tool tailored to the Foundational Phase. Through the elimination of items susceptible to observational bias, the modified formative assessment becomes a concise and objective measurement tool, ready for adoption by educators and future researchers to evaluate children's cognitive science achievements during the Foundational Phase.

Conclusion

The research results indicate that six statement items of the formative assessment instrument for STEAM learning via the volcanic eruption experiment were empirically valid, specifically Items 2, 3, 4, 5, 7, and 10. The initial 10-item instrument demonstrated a Cronbach's Alpha reliability value of 0.752, categorized as high. These findings suggest that the six valid items hold potential for assisting teachers in observing early childhood science process skills during the Foundational Phase. This study is limited by its small sample size of 12 children and its restriction to a single early childhood center. Consequently, the findings cannot be broadly generalized and warrant further testing across more diverse learner characteristics and educational environments.

In practice, early childhood educators can utilize the six valid items as observation guidelines during science experiment activities. Teachers are advised to conduct individual observations, provide children with opportunities to respond independently, and supplement evaluations with anecdotal notes to minimize observational bias. Assessment results can serve as a foundation for planning learning stimulations aligned with children's developmental needs. Future research should involve larger sample sizes and multiple early childhood centers to re-evaluate instrument validity and reliability, thereby establishing stronger empirical evidence regarding instrument consistency and its broader applicability across varied early childhood learning contexts.

References

Anggreni, R. A., Antara, P. A., & Ujianti, P. R. (2022). Instrumen literasi sains pada anak. *Jurnal Pendidikan Anak Usia Dini Undiksha,* 10*(2), 291–301.* https://doi.org/10.23887/paud.v10i2.49303

Akbar, F., Ulumiyah, M., & Wirdah, C. (2024). Implementasi Instrumen Penilaian Pembelajaran pada Anak Usia Dini di RA Sholahiyah Bae Kudus dan

perkembangannya. Hal ini juga membantu dalam memberikan informasi kepada orang. *Jurnal Raudhah*, *12*(1), 67–77 https://jurnaltarbiyah.uinsu.ac.id/index.php/raudhah/article/view/3260

Anggun, Ahirotunisa, Munawaroh, S., & Hidayatu. (2024). Efektivitas Eksperimen Sederhana " Gunung Meletus" untuk Meningkatkan Minat Belajar dan Sikap Sains Anak di TK NU Al-Madani. Motekar : J*urnal Pendidikan Islam Anak Usia Dini*, *1*(1), 1–13.

Bashooir, K., & Supahar. (2018). Validity and reability instrument of scientific literacy performance assesment in physics teaching based on STEM. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *22*(2), 168–181.

Chintya, P., & Andriani, A. (2025). Mempelajari Proses Letusan Gunung Berapi melalui Eksperimen. *Jurnal Ilmiah Research Student*, *2*(2), 643–655.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). SAGE Publications.

Eliamah, E., Wahira, W., & Alam, K. (2022). Meningkatnya Motivasi Belajar Anak Usia Dini (AUD) Melalui Pembelajaran Sains. *EDUSTUDENT: Jurnal Ilmiah Pendidikan Dan Pembelajaran*, *1*(2), 71. https://doi.org/10.26858/edustudent.v1i2.2649 5

Fadhila, I., Azahra, N., & Amiati, D. R. (2025). Mempelajari Proses Erupsi Gunung. *Jurnal Ilmiah Nusantara (JINU)*, *2*(4), 1015–1020.

Nisa, S. T. (2024). Mengenal Gejala Alam Melalui Pendekatan Interaktif Pada Anak Kelompok B Melalui Sains Sederhana Gunung Meletus. *Edukids: Jurnal Pertumbuhan, Perkembangan, Dan Pendidikan Anak Usia Dini*, *21*(1), 158–164. https://doi.org/10.17509/edukids.v21i1.68450

Qonita, Q., Nugraha, D., & Iskandar, B. (2025). Analisis Validitas dan Reliabilitas Instrumen Sikap dan Keyakinan Guru PAUD dalam Pembelajaran STEAM. *Jurnal Obsesi : Jurnal Pendidikan Anak Usia Dini*, *9*(6), 3580–3591.https://doi.org/10.31004/obsesi.v9i6.7714

Sriyono, S. (2021). Metode eksperimen untuk meningkatkan pembelajaran sains pada anak usia dini. *JPI: Jurnal Ilmiah Pendidikan*, *6*(4), 187. https://doi.org/10.20961/jpiuns.v6i4.50068

Subiani, Wulandari, R. S., & Arkam, R. (2022). Peningkatan hasil pembelajaran sains aud melalui metode eksperimen. *Jurnal Mentari*, *2*(1), 45–55.

Syahrul, F. S., Marzuki, K., & Kurnia, R. (2023). Pengaruh Metode Eksperimen Terhadap Keterampilan Proses Sains Anak Usia 5-6 Tahun. *Ihya Ulum: Early Childhood Education*

Journal, 1(3), 123–141. https://doi.org/10.59638/ihyaulum.v1i3.107

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education,* 48, 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Toyyib, M. (2024). Pengenalan Fenomena Alam Melalui Eksperimen Miniatur Gunung Berapi Di TK Plus. *WALADI : Wawasan Belajar Anak Usia Dini. 2*(1), 26–41.