Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Development of a STEM-based argumentation ability assessment instrument for university students within the independent curriculum

Sapuadi Sapuadi [1,*], Zulfa Jamalie [2], Fahmi Hamdi [2], Muhammad Nasir [1]

[1]*Faculty of Tarbiyah and Teacher Training, Universitas Islam Negeri Palangka Raya, Palangka Raya, Indonesia*
[2]*Postgraduate Program, Universitas Islam Negeri Antasari, Banjarmasin, Indonesia*

ARTICLE INFO

Article history:
Received 27 March 2025
Received in revised form
9 July 2025
Accepted 25 July 2025

Keywords:
Argumentation skills
STEM education
Independent curriculum
Instrument development
Rasch model

ABSTRACT

In Indonesia's Independent Curriculum era, universities are expected to help students develop critical thinking and problem-solving skills through interdisciplinary learning. One key skill that students, especially in STEM (Science, Technology, Engineering, and Mathematics) fields, need to develop is argumentation. This includes logical reasoning, using evidence to support claims, and engaging in scientific discussions. However, current assessment tools often do not fully measure students' argumentation skills in STEM areas. This study aims to develop a STEM-based argumentation ability instrument using Item Response Theory (IRT). The content validity of the instrument was examined through a Focus Group Discussion with five experts, and its scoring was reviewed by seven panel members. Construct validity was tested in two stages: a small-scale trial with 15 students and a larger test with 42 students from two classes. The Rasch Model was used to analyze the instrument's validity and reliability. All items fit the Rasch Model, showing that the instrument is both valid and reliable. Therefore, it can be used to assess university students' STEM-based argumentation skills within the Independent Curriculum.

1. Introduction

The learning of the 21st century emphasizes critical thinking and problem-solving skills, creativity and innovation, collaboration, and communication. STEM literacy is an alternative to tackling educational challenges in the 21st century. So, problem-solving using STEM literacy becomes vital in the 21st century (Chu et al., 2021). Critical thinking skills in problem-solving are needed to show reasons, make claims, show evidence, interpret, analyze, and evaluate arguments. Meanwhile, STEM literacy is beneficial for triggering an individual's ability to apply science, technology, engineering, and mathematics concepts to solve problems that cannot be solved using one discipline (Jackson and Mohr-Schroeder, 2018). STEM positively impacted academic achievement and the development of different skills (Batdi et al., 2019). Critical thinking skills in STEM can be developed through argumentation ability.

Argumentation ability refers to the skills and patterns used in constructing and evaluating arguments. Argumentation ability is essential in training students to familiarize themselves with the reasoning behind making decisions based on valid evidence. Involving students in argumentation ability can provide benefits such as helping them: (1) comprehend scientific ideas; (2) acquire 21st century skills; (3) utilize evidence to back up assertions; (4) apply logic; (5) evaluate opposing viewpoints; and (6) comprehend the essence of science. Students' capacity for arguing can improve their conceptual knowledge. They can gain a deeper comprehension of the subject matter by developing their argumentation skills, which involve presenting and defending assertions. Additionally, the ability to argue can help students become critical thinkers of other people's assertions, scientifically literate, and able to articulate their own beliefs with reasoning and supporting data.

Several studies have examined argumentation ability in different contexts. Putra et al. (2023) found that students' argumentation skills in solving statistical problems were influenced by their Adversity Quotient (AQ) levels. Cebrián-Robles et al. (2022) investigated the argumentation ability of pre-

* Corresponding Author.
Email Address: sapuadi@iain-palangkaraya.ac.id (S. Sapuadi)
https://doi.org/10.21833/ijaas.2025.08.016
Corresponding author's ORCID profile:
https://orcid.org/0000-0003-3485-5675

service teachers. They found that while they were proficient in identifying evidence and constructing warrants, they struggled with providing counter-critiques and constructing comparative arguments (Cebrián-Robles et al., 2022). Hasmaningsih et al. (2022) examined the scientific argumentation ability of biology students and discovered that they still needed to improve in both their written and oral argumentation. Aybek (2023) explored using the standard distribution curve to convert the computed item difficulty statistics according to classical test theory (CTT) into the item difficulty parameter of IRT and to assess the efficacy of this transformation using the Rasch model.

Previous research indicates that there is currently no argumentation ability instrument specifically designed for university students within the framework of the Independent Curriculum using a STEM-based context. This gap highlights the need for a comprehensive and validated assessment tool to measure students' argumentation skills in STEM-related learning environments effectively. Such an instrument would provide valuable insights into students' abilities to construct claims, provide evidence, and apply reasoning within real-world STEM contexts. Moreover, it could be helpful for educators to evaluate and enhance students' critical thinking and problem-solving skills, ultimately improving the quality of STEM education in higher institutions. Therefore, further research is necessary to design, develop, and validate a STEM-based argumentation ability instrument tailored to the specific needs of university students in the Independent Curriculum.

The STEM context is critical to training argumentation ability. Thus, developing an argumentation ability measurement instrument based on the STEM context is necessary. Considering the IRT method has many advantages over CTT, this research aims to develop a STEM-based argumentation ability instrument. Therefore, this study aims to develop a valid and reliable STEM-based argumentation ability instrument tailored for university students. By integrating key components of argumentation, such as claim, evidence, reasoning, and rebuttal, this instrument is expected to enhance students' analytical thinking and contribute to the effectiveness of STEM education within the Independent Curriculum framework. The result of the development of this instrument is expected to be used to measure students' argumentation ability based on the STEM context in university students in the Independent Curriculum.

2. Methodology

The development procedure namely tests specifications development, test question writing, test question examination, test trial conduct, test item analysis, test improvement, test assembly, test conduction, and test results interpretation. The development of test specifications is included in the analysis stage. Test question writing, examination, and test trial activities were conducted in the design stage. Activities include analyzing test items, test improvement, and test assembly, including the development stage. Test conduction and test results interpretation activities include the evaluation stage.

The validation of argumentation ability items was initiated by expert and practitioner trials in focus group discussion (FGD) to obtain content validation. This method tests the consensus of experts on the feasibility of the argumentation ability instrument. Three stages of gathering information were needed to reach a consensus. The first stage was to get opinions and recommendations for improving the argumentation ability instrument. The questions used in the FGD were open-ended, allowing participants to provide answers and explanations (Krueger, 2014). In the second stage, the results of the first product improvement were sent to each expert panelist to re-evaluate the products that had been developed. In the third stage, a questionnaire was given to the panelists to assess the content validity of the argumentation ability instrument. Items already good according to content validation are then carried out with a limited-scale trial to get construct validation results.

The number of items on argumentation ability that experts validated was nine questions. The questions represent three claims, three evident, and three reasoning indicators. The number of items on argumentation ability that experts validated was nine questions. The questions represent three claims, three evident, and three reasoning indicators. Questions one to three on argumentation ability were created based on the fiber optic STEM context. Questions four to six are based on the STEM context of endoscopy. Questions seven to nine are based on the STEM context of seismic refraction methods.

Five experts were involved in the FGD process of the argumentation ability instrument, and the Aiken test (Aiken, 1980) involved seven panelists. The number of subjects involved in the limited-scale instrument trial was 15 students. The wide-scale trial involved two classes consisting of 42 students.

Both questionnaire and open-ended data from focus group discussions were subjected to qualitative descriptive analysis. The Aiken Coefficient is cited in the expert consensus criteria for each argumentation ability instrument validity indicator. The content validity coefficient formula proposed by Aiken is as follows.

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

where, $s = r - l_0$, $l_0 =$ lowest validity rating score, $c =$ highest validity rating score, $r =$ the score given by the rater.

The item scoring scale used in this study consisted of 4 scales, in which the range of values given was 1 (lowest) and 4 (highest). The rater consists of 7 experts. Based on these data, n=7, lo=1, and c=4. The instrument content was refined in light of expert feedback. The Aiken coefficient criteria are:

0.0 to 0.4, including low category, or cannot be used for research, 0.4 to 0.8, including medium category, or can be used with improvement, and 0.8 to 1.00, including high category or can be used for research.

The reliability of the validation results of the argumentation ability instrument was based on a statistical analysis of the Percentage of agreement. The validation results of the learning model could be reliable if the reliability value were obtained at 75% (Borich, 2016). The calculation of the reliability of the argumentation ability instrument was strengthened by using Cronbach's Alpha analysis.

Analysis of wide-scale trial data using IRT analysis assisted by the Quest program. The provisions of the threshold value as a reference for determining the difficulty level of the questions are shown in Table 1.

Table 1: Criteria for question difficulty level score

Threshold	Description
$b > 2$	Very difficult
$1 < b \leq 2$	Difficult
$-1 \leq b \leq 1$	Medium
$-1 > b \leq -2$	Easy
$b < -2$	Very easy

The estimation of the test taker's ability to determine the difference can be seen in Table 2. Table 3 displays the item quality criteria based on the Item Response Theory approach.

Table 2: Criteria for differentiating power values

Estimate	Description
$> +1.00$	High ability
-1.00 SD + 1.00	Medium ability
< -1.00	Low ability

Table 3: Quality criteria for items

Criteria	Infit MNSQ	Outfit t
Good	0.90 ≤ infit MNSQ ≤ 1.10	t ≤ 2.00
Quite good	0.77 ≤ infit MNSQ < 0.90 or 1.10 < infit MNSQ ≤ 1.33	t ≤ 2.00
Not good	Infit MNSQ < 0.77 or infit MNSQ > 1.33	t > 2.00

3. Results and discussion

3.1. Analysis

The literature review results at the developing test specifications stage show that indicators of argumentation ability are in Table 4.

Table 4: Indicators of argumentation ability according to experts

Reference	Indicator	Description
Sidorova et al. (2023)	Claim	Ability to state statements
	Evidence	Ability to show evidence
	Reasoning	Ability to link claims and evidence
	Rebuttal	Ability to dispute the veracity of claims
Sandoval (2014)	Articulation	Ability to claim
	Warrant	Ability to show data to support claims
Li et al. (2018)	Claim	Ability to state the relationship between independent and dependent variables
	Evidence	Ability to present sufficient and precise data
	Reasoning	Ability to theorize, relate data and claims, and support or refute claims

An indicator of argumentation ability is an argument obtained from a phenomenon, using relevant evidence and reason to support the argument. Argumentation ability refers to how students can articulate claims causally and whether their claims are warranted by the data they examined during the investigation (Sandoval, 2014). The argumentation ability referred to by the researcher is the ability to explain the reasons for the relationship between claims and evidence and convince of the truth of a reason.

In general, indicators of argumentation ability are derived from the aspect of Toulmin's argument (Magalhães, 2020). Toulmin's argument pattern framework includes six aspects of the argument, namely: (1) statement (claim), (2) evidence (evidence), (3) justification (warrant), (4) support (backing), (5) qualification (qualifier), and (6) rebuttal. Claims result from established values, opinions about the existing situation, and affirmations of points of view. Evidence is facts that are used to support a claim. Justification is the reason that links the data to the claim. Support is a basic assumption in a particular field that supports justification. Qualification is a situation where the claim is accurate. Disclaimers are cases where claims are untrue or unsupported by data, justification, and support.

3.2. Design

At the examination test questions stage, the researcher made a prototype of an argumentation ability using a grid of argumentation ability instruments. Researchers used the argumentation ability indicators, according to the experts in Table 4, as a reference to determine the indicators to be measured in this study. The indicators to be measured in this study are shown in Table 5.

Table 5: Argumentation ability indicators

Indicator	Description
Claim	Ability to provide an assessment of a statement
Evidence	Ability to show evidence in the form of facts/concepts/laws/principles/or supporting claims
Reasoning	Ability to provide logical reasoning relationships between claims and evidence

A grid of argumentation ability instruments is arranged as in Table 6, based on the construction of argumentation ability indicators according to researchers in Table 5.

3.3. Develop

Argumentation ability questions are arranged based on a predetermined grid. The prototypes of the argumentation ability items that have been prepared are then continued with content and construct validation activities. Table 7 shows suggestions for improvement and follow-up to recommendations for improving the prototype of argumentation ability instruments.

The content validity of the argumentation ability instrument was tested using the Aiken coefficient. The results of the Aiken coefficient analysis can be seen in Table 8.

As can be seen in Table 8, every argumentation ability item satisfies content validity and reliability requirements because all of the Aiken coefficients fall into the high category. Content validity reflects that the argumentation ability instrument has a state-of-the-art approach based on a clear theory. The reliability test results of argumentation ability obtained Cronbach's Alpha (α)=0.77. The argumentation ability instrument is categorized as very reliable because the calculated value of Cronbach's Alpha (α) is more than 0.7.

Table 6: Grid of argumentation ability questions

Indicator	Question indicator	Question number
Claim	Student submits a claim statement about the STEM context.	1a, 2a, and 3a
Evidence	Students submit sufficient evidence in the form of facts/concepts/or laws that support claims.	1b, 2b, and 3b
Reasoning	Students explain the logical reasoning of the relationship between claims and evidence.	1c, 2c, and 3c

Table 7: Suggestions and follow-ups for improvement of the argumentation ability instrument

No.	Aspect	Improvement suggestions	Follow-up activities
1	Indicator formula	Indicators of competency achievement should be formulated referring to the audience, behavior, condition, and degree formula.	Develop indicator questions using the audience, behavior, condition, and degree formula.
2	Assessment rubric	The scoring rubric should not be too specific on the concept, but in the form of possible answer criteria.	Develop an assessment rubric based on the answer criteria.
3	Context representation	Clear and relevant images should represent the context of the question.	Presents a clear and relevant image in the context of the question.
4	Question structure	Questions are made open-ended, which allows more than one answer.	Making questions open-ended to stimulate different types of answers.

Table 8: The result of the Aiken coefficient

Item	Σ (Sum)	V (Aiken's V)	Criteria	Decision	R (%)	Reliability
1	21	1.05	High	Valid	86%	Reliable
2	21	1.05	High	Valid	86%	Reliable
3	21	1.05	High	Valid	86%	Reliable
4	19	0.95	High	Valid	86%	Reliable
5	19	0.95	High	Valid	86%	Reliable
6	19	0.95	High	Valid	86%	Reliable
7	21	1.05	High	Valid	100%	Reliable
8	19	0.95	High	Valid	100%	Reliable
9	21	1.05	High	Valid	100%	Reliable

3.4. Evaluate

After the argumentation ability items meet the content validity, a limited-scale trial is conducted, which aims to test construct validity. The construct validity analysis used the IRT method. The IRT analysis reveals the item difficulty and test takers' ability to respond to questions, which is impossible with traditional test analysis. Fig. 1 displays the test participants' ability to distinguish between items and the degree of difficulty.

The benefit of IRT analysis is that, as Fig. 1 illustrates, it can depict the distribution of items' suitability for the Rasch model. Items that "fit" the Rasch model refer to items that meet its basic assumptions. Since every item in Fig. 2 falls between 0.77 and 1.33, all questions fit the Rasch model. Table 9 provides a concise summary of the items' quality. According to the Rasch model, which is based on item response theory, an individual's likelihood of correctly answering a question item is solely determined by their aptitude and the item's difficulty, with no influence from outside sources. Items fitting the Rasch model have the following meaning: (1) They exhibit invariance about population characteristics. (2) The Rasch model assumes that an individual's ability level is the only factor influencing the likelihood of a correct response. (3) Items that fit the Rasch model have high reliability and validity according to the measured construct. This means the items provide good information about the respondent's ability level.

Table 9 shows that all argumentation ability questions are in a good category. Thus, all items of argumentation ability meet construct validity. Construct validity reflects that the components of the argumentation ability instrument are consistently developed. Because the argumentation ability instrument is valid and reliable. It can be

implemented to measure the ability of students to argue. The item difficulty level, or threshold, suggests that all questions fall into the medium category, except item number 3, which is in the easy category, all the questions that test takers with high and moderate proficiency can answer. Only question number 3 is one that test takers with low ability can answer. The difficulty level of a question is an important parameter in the context of IRT. The degree of difficulty measures how difficult or easy a question is for the respondent being tested. The degree of difficulty of the questions has several implications and benefits in the context of measurement and evaluation. Firstly, the degree of difficulty helps determine the ability required to answer the question correctly. Questions with a high degree of difficulty require a higher ability to be answered correctly.

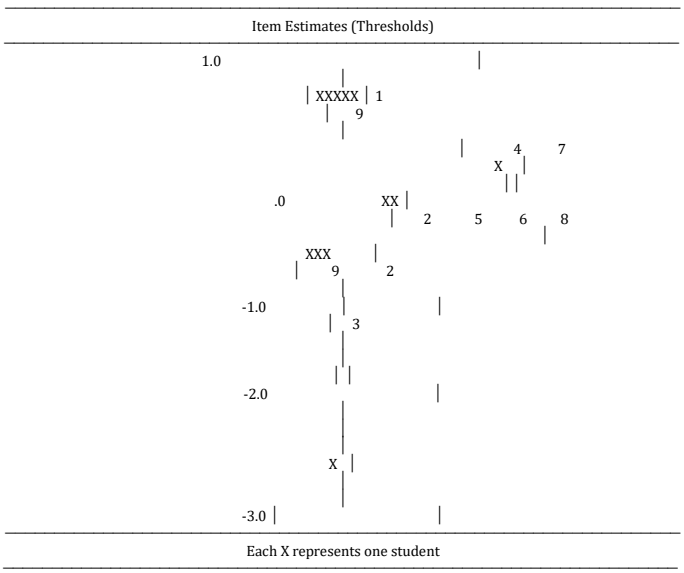

Fig. 1: Level of difficulty and distinguishing ability of the test participants

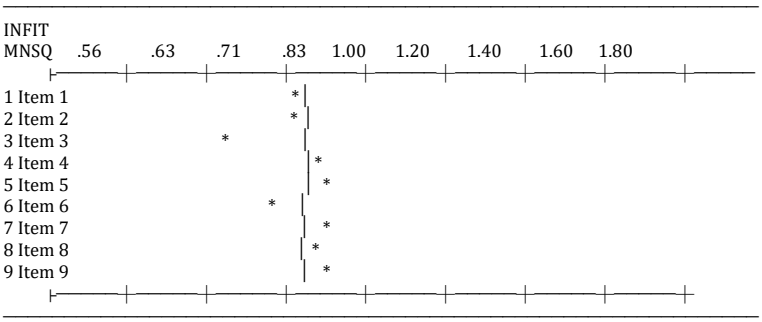

Fig. 2: Distribution of item suitability with the Rasch model

Table 9: Recapitulation of the quality of argumentation ability items

Item No.	Infit MNSQ	Outfit t	Difficulty parameter (B)	Quality
1	.98	-.1	.85	Good
2	.95	-.1	-.20	Good
3	.78	-.5	-1.01	Good
4	1.03	.1	0.49	Good
5	1.11	.4	-.20	Good
6	.87	-.3	-.20	Good
7	1.12	.3	.49	Good
8	1.03	.1	-.20	Good
9	1.12	.7	-.20	Good

In contrast, individuals with a lower ability level can answer questions with low difficulty. Secondly, by understanding the difficulty of each question, test developers can design tests that suit the measurement objectives. Too easy or too complicated questions may not provide good information about an individual's abilities.

Differential item functioning on a question refers to the extent to which different levels of ability or group characteristics can influence how respondents answer the question. The differentiation power function has several implications and benefits in the context of tests and measurements. Firstly, Discrimination helps identify whether a question shows bias towards a particular group. Question bias occurs when groups with the same ability level answer a question differently. Secondly, discriminability can help investigate whether a test can be considered equivalent between different groups. If there is significant differential power, this

may indicate that the test is not entirely equivalent across groups. Additionally, by paying attention to differential power, the interpretation of test results can be improved to be more accurate and objective, especially if the test is used to make important decisions, such as selection or assessment.

An example of a valid and reliable STEM-based argumentation ability question is in Fig. 3.

> Advances in medical science and biotechnology raise new questions in Islamic jurisprudence, such as the rulings on organ transplantation, vaccination, and the use of DNA in identification. Use the STEM approach and Islamic evidence to develop an argument about how Muslims should respond to modern medical technology. Is the use of biotechnology justified according to Islamic law? Explain with examples and relevant scientific and Islamic bases!

Fig. 3: An example of a STEM-based argumentation ability question

The argumentation ability instrument has met both content validity and construct validity. The content validity of the argumentation ability instrument is shown by 100% of the Aiken coefficients for the items in the argumentation ability instrument with valid categories. The construct validity of the argumentation ability instrument was shown by 100% of the results of the IRT test items with good categories. The argumentation ability instrument is valid, meaning that the instrument is of state-of-the-art quality, has a strong theoretical and empirical basis, and has consistency between model indicators.

The validity of argumentation ability instruments has been examined in several studies. Nuzulah et al. (2023) developed an inquiry-based learning tool to assess students' scientific argumentation abilities, which was found to have high validity and reliability. Ariawan et al. (2022) developed an instrument to assess mathematical critical thinking skills, which was also found to have high validity. Yonata et al. (2022) developed an answer argumentation instrument to identify students' misconceptions, which was validated for content and construct validity. Affandy et al. (2021) calibrated an instrument for argumentation skills in Fluid Statics using item response theory, and the instrument was found to be valid and reliable. These studies demonstrate the importance of validating argumentation ability instruments to ensure accuracy and effectiveness in assessing students' skills and knowledge.

Certain traits distinguish scientific argumentation ability: (1) it typically aims to understand phenomena by taking into account other scientific facts or formulating new theories to explain the behavior of new phenomena; (2) it tends to be more systematic, deeper, and more accurate than arguments based solely on common sense. The use of relevant and sufficient evidence in this instance and the persuasiveness of the argument supporting the evidence are considered when evaluating the argumentation. Four components are essential to a strong argumentative ability: Causality, conceptual framework, relevance, and a suitable degree of representation (de Andrade et al., 2019). A strong argumentative ability requires that the data be pertinent to the phenomenon. The argumentation ought to offer a theoretical framework grounded in scientific theories. Arguments must follow a consistent, logical causal narrative linking a phenomenon to many underlying processes.

The argumentation ability instrument meets the reliability indicated by the Percentage of agreement value for each indicator is more than 75%. Overall, the reliability test results of the argumentation ability instrument are reliable. The level of reliability is indicated by the value of Cronbach's Alpha (α) = 0.77, with the category of excellent reliability. The reliable argumentation ability can then be implemented to measure students' argumentation ability.

Multiple studies assessed the reliability of the argumentation ability instruments. The instrument developed for argumentation skills on the subject of Fluid statistics showed high reliability, with a reliability value of 0.86 (Affandy et al., 2021). In a study on inquiry-based learning tools, the validity index for the tools assessing students' scientific argumentation abilities was high, indicating high reliability (Bisra et al., 2018). These studies demonstrate that the argumentation ability instruments developed and used in various contexts are reliable.

Benefits of IRT compared to CTT. Firstly, IRT allows for a causal interpretation of the latent scores, providing a deeper understanding of the underlying causes of ratings (Veldkamp et al., 2024). Secondly, IRT enables the evaluation of an item's capacity to distinguish between people with high and low levels of the attribute it is meant to measure. CTT does not offer this degree of item-specific data. Additionally, IRT is more robust in handling missing data. CTT often requires complete data for each individual. At the same time, IRT can still provide estimates even if some items are missing, given that the missingness is unrelated to the measured trait. Furthermore, IRT is based on a more solid theoretical foundation, involving probabilistic models that describe the relationship between a person's ability and the probability of a correct response. This makes IRT more consistent with modern measurement theory.

Each argumentation indicator can be functionally linked to STEM competencies in STEM-based learning as presented in Fig. 4.

Even though the researchers tried to control for bias in this research, there were still limitations. These limitations include indicators of argumentation ability that only use three aspects: claim, evidence, and reasoning. It is hoped that future researchers can develop argumentation ability instruments with more indicators, such as support (backing), qualification (qualifier), and rebuttal aspects. Science education practitioners can

use the instruments developed to measure STEM-based argumentation abilities in STEM-based learning.

Critically, the success of Rasch modelling shows that the instrument can capture the variation in respondents' abilities objectively and linearly, and minimize bias that may arise from unequal distribution of responses or item mismatch. However, although these results are encouraging, it is important to consider several limitations. First, content validity, measured through item fit in the Rasch model, does not yet include overall construct validity, which should also be supported by confirmatory factor analysis (CFA) or external validity through correlation with related variables. Second, the generalizability of these findings is still limited to the context of the participants involved in the study, so cross-population or institutional testing is needed to ensure the instrument's reliability more broadly.

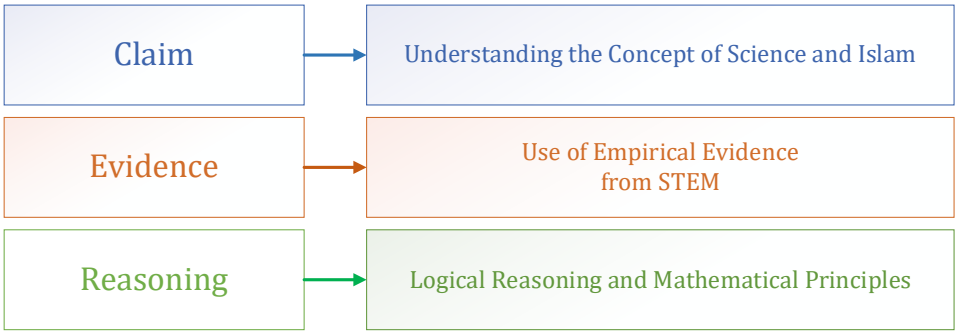

Fig. 4: Relationship between argumentation ability indicators and STEM competencies

When compared to previous literature, these results are in line with findings stating that argumentation-based instruments in the context of science and STEM must demonstrate structural fit (Wisutama et al., 2023; Gu, 2021) and internal consistency (Yovita et al., 2024; Mao et al., 2018) in order to be used for both formative and summative assessments. While the focus on structural fit and internal consistency is critical, some argue that the dynamic nature of classroom assessments may require flexibility in these criteria to adapt to diverse learning environments and student needs. This perspective suggests that rigid adherence to structural norms might overlook the nuanced realities of formative assessment practices. This study extends these contributions by adapting the argumentation assessment framework to the context of the Merdeka Curriculum, which emphasizes differentiated and project-based learning.

Furthermore, the validity and reliability of this instrument can support the development of authentic assessments in higher Education, especially in strengthening critical thinking, collaborative, and complex problem-solving skills that are the main characteristics of STEM-based learning (Qudratuddarsi et al., 2022; Chusni and Suherman, 2021). STEM-focused assessments encourage the development of higher-order thinking skills, vital for solving complex problems in real-world contexts (Sari et al., 2023; Setyawati et al., 2023). Further research is recommended to evaluate the instrument's sensitivity in detecting changes in students' argumentative abilities before and after learning interventions and to explore the potential for digitizing the instrument to make it more applicable in online learning.

Implementing STEM-Based Argumentation Skills, Instruments can be used with rubric guidance. This assessment tool has been developed and validated to reflect the quality of students' arguments in a purposeful manner. In the context of STEM, this rubric usually includes aspects such as clarity of claims, strength of evidence, the relationship between claims and evidence, and logic or scientific reasoning. The use of rubrics ensures consistency of assessment and makes it easier for educators to identify the strengths and weaknesses of students' arguments. The standardized rubric for assessing STEM-based argumentation skills is shown in Table 10.

Table 10: STEM-based argumentation skills rubric

Rated aspect	Score 3 (very good)	Score 2 (enough)	Score 1 (less)
Claim	Claims are explicit, relevant to the topic, and stated precisely.	The claim is clear enough, but lacks specificity or is vague.	Claims are unclear, irrelevant, or non-existent.
Evidence	Strong, relevant evidence based on valid scientific data/information.	Evidence is presented, but it is not strong or relevant.	The evidence is irrelevant, weak, or non-existent.
Reasoning	A logical and scientific explanation that links evidence and claims well.	Explanations exist, but lack logic or do not fully connect the evidence to the claim.	There is no explanation or logic in connecting the evidence to the claim.
Use of STEM concepts	STEM concepts are used accurately to support the argument.	STEM concepts are used, but errors or a lack of depth exist.	Failure to use or misuse STEM concepts.

Data from the assessment results are used to diagnose students' difficulties in arguing and to design further learning steps. For example, if many students fail to connect evidence to claims, teachers can design special learning sessions on "reasoning" or provide examples of good arguments. This ensures the learning process is adaptive and responsive to students' needs.

4. Conclusion

All the questions fall into a good category according to content and construct validity using Item Response Theory, which is based on the Rasch model. The argumentation ability instrument's state-of-the-art, which is founded on a well-defined theory, is reflected in its content validity. The argumentation ability instrument's construct validity indicates that each part was created consistently. Cronbach's Alpha (α=0,77) is calculated, and since it is greater than 0.7, the argumentation ability instrument is classified as very reliable. The argumentation ability instrument can be used to assess students' STEM-based argumentation abilities because it is a valid and reliable instrument.

This research's limitation is that the scope of developing argumentation ability is only to claim, evidence, and reasoning indicators. Thus, it is recommended that further research develop the same instrument but with indicators that involve support (backing), qualification (qualifier), and rebuttal aspects. In addition, practitioners can use this instrument to measure students' argumentation abilities in STEM-based learning.

Future research is recommended to develop a more comprehensive instrument that includes the indicators of claim, evidence, and reasoning and support (backing), qualification (qualifier), and rebuttal to provide a more holistic assessment of students' argumentation skills. Additionally, the instrument should be tested for validity and reliability to ensure its accuracy and applicability in research and educational practice. Practitioners can utilize this instrument to assess and enhance students' argumentation abilities in STEM-based learning, helping educators design more effective pedagogical interventions. Furthermore, developing a user guide for educators, including practical examples and assessment methods, is essential to facilitate its implementation in various learning contexts.

List of abbreviations

AQ	Adversity quotient
CFA	Confirmatory factor analysis
CTT	Classical test theory
DMU	Decision-making unit
FGD	Focus group discussion
IRT	Item response theory
MNSQ	Mean square (fit statistics)
SD	Standard deviation
STEM	Science, technology, engineering, and mathematics

Acknowledgment

This research was supported by funding from IAIN Palangka Raya. The authors gratefully acknowledge the IAIN Palangka Raya Research and Community Service.

Compliance with ethical standards

Ethical considerations

Participation was voluntary, and informed consent was obtained from all experts and students involved. All data were anonymized to protect participants' privacy.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Affandy H, Nugraha DA, Pratiwi SN, and Cari C (2021). Calibration for instrument argumentation skills on the subject of fluid statics using item response theory. Journal of Physics: Conference Series, 1842: 012032. https://doi.org/10.1088/1742-6596/1842/1/012032

Aiken LR (1980). Content validity and reliability of single items or questionnaires. Educational and Psychological Measurement, 40(4): 955-959. https://doi.org/10.1177/001316448004000419

Ariawan R, Nurmaliza N, Dahlia A, Nufus H, and Nurdin E (2022). Validity of mathematical critical thinking ability assessment instruments. Jurnal Cendekia: Jurnal Pendidikan Matematika, 6(3): 2673-2684. https://doi.org/10.31004/cendekia.v6i3.1636

Aybek EC (2023). The relation of item difficulty between classical test theory and item response theory: Computerized adaptive test perspective. Journal of Measurement and Evaluation in Education and Psychology, 14(2): 118-127. https://doi.org/10.21031/epod.1209284

Batdi V, Talan T, and Semerci Ç (2019). Meta-analytic and meta-thematic analysis of STEM education. International Journal of Education in Mathematics, Science and Technology, 7(4): 382-399.

Bisra K, Liu Q, Nesbit JC, Salimi F, and Winne PH (2018). Inducing self-explanation: A meta-analysis. Educational Psychology Review, 30: 703-725. https://doi.org/10.1007/s10648-018-9434-x

Borich GD (2016). Observation skills for effective teaching: Research-based practice. Routledge, London, UK. https://doi.org/10.4324/9781315633206

Cebrián-Robles D, Osorio JH, Mariscal AJF, and Lorite IMC (2022). Assessing the argumentation ability of pre-service teachers: Case study concerning the chemical dissolution process. International Journal of Educational Research and Innovation, 17: 73-83. https://doi.org/10.46661/ijeri.4968

Chu SKW, Reynolds RB, Tavares NJ, Notari M, and Lee CWY (2021). 21st century skills development through inquiry-based learning from theory to practice. Springer, Singapore, Singapore. https://doi.org/10.1007/978-981-10-2481-8

Chusni MM and Suherman S (2021). Developing authentic assessment instrument based on multiple representations to measure students' critical thinking skills. Momentum: Physics Education Journal, 5(2): 194-208. https://doi.org/10.21067/mpej.v5i2.5790

de Andrade V, Freire S, and Baptista M (2019). Constructing scientific explanations: A system of analysis for students' explanations. Research in Science Education, 49: 787-807. https://doi.org/10.1007/s11165-017-9648-9

Gu PY (2021). An argument-based framework for validating formative assessment in the classroom. Frontiers in

Education, 6: 605999. https://doi.org/10.3389/feduc.2021.605999

Hasmaningsih L, Karnan K, and Handayani BS (2022). Level of scientific argumentation ability of students in biology learning. Jurnal Pijar Mipa, 17(6): 717-722. https://doi.org/10.29303/jpm.v17i6.4242

Jackson CD and Mohr-Schroeder MJ (2018). Increasing STEM literacy via an informal learning environment. Journal of STEM Teacher Education, 53(1): 4. https://doi.org/10.30707/JSTE53.1Jackson

Krueger RA (2014). Focus groups: A practical guide for applied research. SAGE Publications, New York, USA.

Li Z, Oren N, and Parsons S (2018). On the links between argumentation-based reasoning and nonmonotonic reasoning. In: Black E, Modgil S, and Oren N (Eds.), Theory and applications of formal argumentation. TAFA 2017. Lecture Notes in Computer Science, Vol. 10757. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-319-75553-3_5

Magalhães AL (2020). Teaching how to develop an argument using the Toulmin model. International Journal of Multidisciplinary and Current Educational Research, 2(3): 1-7. https://doi.org/10.14689/issn.2148-624.1.7c.3s.10m

Mao L, Liu OL, Roohr K, Belur V, Mulholland M, Lee HS, and Pallant A (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. Educational Assessment, 23(2): 121-138. https://doi.org/10.1080/10627197.2018.1427570

Nuzulah DF, Kirana T, and Ibrahim M (2023). Validity of inquiry-based learning tools on students' scientific argumentation ability. International Journal of Recent Educational Research, 4(2): 137-148. https://doi.org/10.46245/ijorer.v4i2.309

Putra ZRIA, Rahardi R, Sisworo S, and Permadi H (2023). Profile of students' argumentation ability based on adversity quotient in statistical problem. Journal of Medives: Journal of Mathematics Education IKIP Veteran Semarang, 7(1): 106-116. https://doi.org/10.31331/medivesveteran.v7i1.2330

Qudratuddarsi H, Hidayat R, Nasir N, Imami MKW, and bin Mat Nor R (2022). Rasch validation of instrument measuring Gen-Z science, technology, engineering, and mathematics (STEM) application in teaching during the pandemic. International Journal of Learning, Teaching and Educational Research, 21(6): 104-121. https://doi.org/10.26803/ijlter.21.6.7

Sandoval W (2014). Science education's need for a theory of epistemological development. Science Education, 98(3): 383-387. https://doi.org/10.1002/sce.21107

Sari DS, Widiyawati Y, Nurwahidah I, and Setiawan T (2023). STEM critical thinking assessment for measuring students' critical thinking skills in the automotive chemistry course. Jurnal Penelitian Pendidikan IPA, 9(7): 5289-5295. https://doi.org/10.29303/jppipa.v9i7.4750

Setyawati RD, Prasad B, Pramasdyahsari AS, and Aini SN (2023). Construct the validity of STEM and project-based critical thinking skills test instruments using the Rasch model. Phenomenon: Jurnal Pendidikan MIPA, 13(1): 96-110. https://doi.org/10.21580/phen.2023.13.1.16246

Sidorova EA, Akhmadeeva IR, Kononenko IS, and Chagina PM (2023). Argument extraction based on the indicator approach. Pattern Recognition and Image Analysis, 33(3): 498-505. https://doi.org/10.1134/S1054661823030410

Veldkamp K, Grasman R, and Molenaar D (2024). Recommendation with item response theory. Behaviormetrika. https://doi.org/10.1007/s41237-024-00244-3

Wisutama RA, Sulaeman NF, and Zulkarnaen Z (2023). Argumentation skill in STEM-EDP worksheet for high school students: Validity aspect. Jurnal Pembelajaran Fisika, 12(3): 137-145. https://doi.org/10.19184/jpf.v12i3.42638

Yonata B, Suyono S, and Azizah U (2022). Answers argumentation instrument to strengthen conception diagnostic test on the concept of chemical kinetics: Validity aspect. SHS Web of Conferences, 149: 01007. https://doi.org/10.1051/shsconf/202214901007

Yovita Y, Sonia G, Vebrianto R, Susanti E, and Berlian M (2024). Development of scientific argumentation test instruments for students on the classification of living creatures in junior high school. Science Education and Application Journal, 6(2): 155-164. https://doi.org/10.30736/seaj.v6i2.1056