Development of Assessment as Learning Instruments to Improve Chemistry Pre-Service Teachers' Argumentation Skills Using AI-Assisted Collaborative Argumentation Online Whiteboard

Aulia Rochmah [1], Suyono Suyono [1], Suyatno Sutoyo [1]

[1] *Department of Chemistry Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Jl. Komplek Universitas Negeri Surabaya Gedung D1, Surabaya, Jawa Timur, Indonesia*

Abstract **– The study developed an assessment as learning instrument to foster the argumentation skills of pre-service chemistry teachers. The instrument was designed as an alternative assessment tool that integrates self-assessment, peer assessment, and collaborative discussions on an AI-assisted online whiteboard aligned with Toulmin's argumentation framework. The ADDIE model guided the development of the instrument. The instrument underwent expert content and construct validation, achieving high validity ratings. Pilot testing with 31 participants demonstrated moderate inter-rater reliability (ICC = 0.621) for peer assessments and moderate correlations between self, peer, and instructor assessments, highlighting the need for improved feedback mechanisms and self-assessment training. The analysis of collaborative argumentation discussion revealed that most arguments demonstrated moderate quality, and a smaller proportion achieved the highest quality level. The analysis of practicality and effectiveness gave moderate ratings. Feedback expressed a need for more explicit instructions, more task time, and improved feedback mechanisms. Assessment as learning based instrument was moderately effective but demonstrated some potential for encouraging reflective, collaborative, and logical argumentation processes. Further refinements are needed to optimize its impact.**

Keywords **– Assessment as learning, self-assessment, peer assessment, collaborative argumentation.**

DOI: 10.18421/TEM144-52
https://doi.org/10.18421/TEM144-52

Corresponding author: Suyono Suyono,
Department of Chemistry Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Jl. Komplek Universitas Negeri Surabaya Gedung D1, Surabaya, Jawa Timur, Indonesia
Email: suyono@unesa.ac.id

Received: 06 January 2025.
Revised: 21 July 2025.
Accepted: 05 August 2025.
Published: 27 November 2025.

1. Introduction

Argumentation skill (AS) involves critically evaluating and constructing arguments, enabling individuals to oppose or support claims effectively. AS is not only about conveying opinions clearly and convincingly [1] but also encompasses critical thinking [2], collaboration [3], [4], integrating multiple perspectives [5], and navigating complexity and ambiguity [6]. These skills involve developing, interpreting, and evaluating arguments by identifying whether an argument adheres to deductive or non-deductive inference patterns [7]. For pre-service chemistry teachers, mastering AS is essential in building a solid framework of logical thinking, which forms the foundation for effective teaching and critical analysis [8]. Thus, improving AS can help them become better educators while fostering their critical and analytical thinking abilities.

In the context of chemistry education, AS focuses on two key aspects. First, "learning to argue", which emphasizes developing skills such as constructing valid argument structures, counterarguments, and rebuttals supported by appropriate evidence [9], [10], [11]. Second, "arguing for learning", which highlights the role of argumentation in conceptual change through constructive and dialogic interactions [12], [13], [14].

By engaging in argumentation, students can apply knowledge in context, broaden their perspectives, and address misconceptions [15], [16]. Implementing practical argumentation activities in chemistry learning reduces misconceptions and enhances students' understanding.

AS also plays an important role in developing students' social and collaborative skills. Through collaborative argumentation, students refine their conceptual understanding by working with peers to challenge and validate assumptions [9]. However, research indicates that students' AS is generally at lower levels, with many struggling to incorporate key components such as a warrant, backing, qualifier, and rebuttal into their arguments [17], [18]. These findings highlight the need for effective strategies to improve AS in educational contexts.

To address these challenges, developing Assessment as Learning (AaL) instruments have emerged as a promising solution. AaL incorporates peer and self-assessment, encouraging students to be responsible for their learning, which has been shown to enhance academic success [19]. By using AaL, students can actively reflect on and evaluate their arguments, fostering the development of strong AS [20]. Advances in digital technology, particularly those supported by Artificial Intelligence (AI), open new opportunities to facilitate collaborative AS learning [4], [21].

However, the lack of appropriate assessment instruments tailored to evaluate AS in this context remains a significant challenge.

The present study addresses these gaps by developing an innovative AaL tool for improving the AS of pre-service chemistry teachers. The proposed instrument integrates a Collaborative Argumentation Online Whiteboard, augmented by AI capabilities, to facilitate real-time discourse, joint argumentation, and feedback [4], [21]. Using AI-based tools allows students to recognize the strengths and weaknesses of their arguments, thus fostering metacognitive skills critical to the development of AS [22], [23]. While previous studies have focused on formative Assessment for Learning (AfL) instruments [24], [25], no research has specifically developed AaL instruments for improving AS in pre-service chemistry teachers. Figure 1 shows how students' argumentation skills are improving in this study.

Thus, this study contributes to the field by developing a novel AaL instrument and exploring its feasibility in improving AS. The findings will provide valuable insights for curriculum development and innovative practices in teacher education, equipping future chemistry educators with the skills needed to excel in their profession.

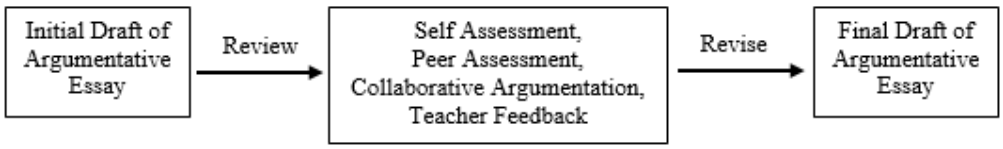

Figure 1. Assessment as learning for students' argumentation skill

2. Methods

This study employed a development research approach with the primary aim of creating the Chemistry Argumentation Skills Assessment as Learning (CASAL) instrument to improve the argumentation skills of pre-service chemistry teachers. The ADDIE development model was adopted, encompassing five stages: Analyse, Design, Develop, Implement, and Evaluate [26]. The feasibility of the CASAL instrument was assessed.

The Analyse stage included competency analysis, curriculum analysis, literature reviews, and evaluations. Competency analysis examined students' profiles and chemistry learning related to AS. Curriculum analysis identified chemical concepts that could be integrated with AS indicators. Literature reviews explored the challenges of AS assessment, empirical research, and theoretical frameworks supporting instrument development.

This stage concluded with evaluations to refine findings, identifying research gaps and problems requiring solutions.

The Design stage focused on developing CASAL criteria and rubrics for AS, incorporating e-questionnaires for self-assessment and peer assessment, and designing an online whiteboard template for collaborative argumentation using Padlet, aligned with Toulmin's AS framework indicators. Validation sheets and student response questionnaires were also prepared. A Focus Group Discussion (FGD) with the research team and five senior lecturers evaluated the instrument's content and structure, leading to revisions. Outputs included finalized rubrics, Padlet templates, validation sheets, and questionnaires.

During the development phase, three content and language experts reviewed the first draft of the CASAL instrument.

Feedback was analyzed descriptively, leading to revisions and the production of a second draft. Validation testing of the second draft involved five field experts using a 4-point Likert scale to measure validity [27]. A minimum mode score of 3 was required to declare validity. Revisions were made based on suggestions, resulting in a third draft deemed valid in construct and content.

The Implementation stage involved pilot testing the third draft with 31 pre-service chemistry teachers. The CASAL instrument included five Student Activity Sheets (SAS). Students drafted argumentative essays (SAS 1), completed self-assessments (SAS 2), conducted peer assessments in groups (SAS 3), engaged in collaborative discussions on Padlet (SAS 4), and wrote final argumentative essays (SAS 5). The reliability of peer assessment was assessed using the Intraclass Correlation Coefficient (ICC) with a two-way random effects model [28]. Practicality was measured via closed and open-ended questionnaires, and effectiveness was evaluated based on students' argumentation activities. Spearman correlation tests assessed the consistency of self, peer, and instructor assessments. Transcripts of collaborative argument discussions on Padlet were coded using Toulmin's indicators, while argument quality was assessed with a five-level model [29], adapted from face-to-face discourse to analyze asynchronous interactions. A rebuttal, central to the highest levels, was identified when a student explicitly referenced a prior claim and challenged its data, warrant, or backing.

3. Results and Discussion

The CASAL instrument developed in this study comprises a collection of documents that evaluate pre-service chemistry teachers' AS through self-assessment, peer assessment, instructor assessment, and collaborative discussions. The instrument aims to enhance AS in chemistry learning by meeting three feasibility criteria: validity, practicality, and effectiveness [30]. The CASAL instrument supports sustainable student learning by fostering reflective, collaborative processes and active involvement in self-assessment.

3.1. Validity

Five experts validated The CASAL instrument, evaluating both content and construct validity. Content validity ensured alignment with accurate scientific knowledge, particularly in chemical kinetics. As seen in Table 1, the results confirmed a highly valid rating across all aspects, free from errors in facts, concepts, principles, or theories of chemical kinetics.

Construct validity confirmed the instrument logical alignment with AS indicators, with a rating mode of 4 (highly valid) across all aspects (Table 2). These findings demonstrate that the CASAL instrument effectively measures the intended constructs.

Table 1. CASAL instrument content validity

Assessment Aspects	Mo	Category
CASAL instrument contains concepts, facts, concepts, principles, laws, and theories that apply to chemical kinetics and are correct and by the rules of chemistry or do not contain misconceptions.	4	Highly Valid
The content of chemical kinetics is relevant as a vehicle for practising student's AS.	4	Highly Valid
The assessment rubric includes all components of Toulmin's argumentation (claim, evidence, warrant, backing, rebuttal, qualifier).	4	Highly Valid

Table 2. CASAL instrument construct validity

Indicators	Assessment Aspects	Mo	Category
Claim	1. Students provide a main statement or claim and relate it to the factors influencing their claim.	4	Highly Valid
	2. Students evaluate whether the reader can understand the claim they made.	4	Highly Valid
Data	1. Students use data from experiments or literature as supporting evidence for their claim.	4	Highly Valid
	2. Students evaluate how well the evidence provided supports and strengthens their claims.	4	Highly Valid
Warrant	1. Students provide a logical explanation that connects the evidence to the claim.	4	Highly Valid
	2. Students evaluate whether the explanation provided ensures coherence between the evidence and the claim.	4	Highly Valid
Backing	1. Students provide additional explanations to strengthen their argument further.	4	Highly Valid
	2. Students reflect on how this explanation reinforces their argument.	4	Highly Valid
Rebuttal	1. Students present counterarguments and refutations against their proposed claims and evidence.	4	Highly Valid
	2. Students evaluate how the counterargument strengthens their argument.	4	Highly Valid
Qualifier	1. Students state the limitations of their claim based on analysis.	4	Highly Valid
	2. Students assess the accuracy of these limitations relative to their claim and evidence.	4	Highly Valid

3.2. Implementation and Reliability

Students drafted essays on reaction rate factors affecting Octobot performance. The initial drafts underwent self-assessment and peer evaluations using a 4-point rubric, as seen in Table 3.

AI detection tools were used to guarantee the originality of the content, meaning some students were requested to revise their essays containing AI-generated content. Research [31] demonstrated the efficacy of these tools. The Isgen.ai platform was selected due to its multilingual detection capabilities, including Bahasa Indonesia content, and its 96.4% accuracy in detecting content produced by AI models such as ChatGPT, GPT4, Bard, LLaMa, and others. A manual cross-checking process was implemented to improve accuracy. The distinctive features of AI-generated essays, which differentiate them from human-generated essays, were considered during this process. One notable distinction is the tendency of AI-generated essays to exhibit neat and consistent sentence structures characterized by uniform patterns in contrast to the more varied and flexible nature of human writing [32]. Additionally, AI-generated essays frequently produce superficial analyses and use repetitive vocabulary. In contrast, human writing exhibit in-depth arguments supported by more contextualized and varied word choices [33].

The transitions between paragraphs in AI essays have been observed to lack natural flow compared to human writing, where such transitions are organic. Additionally, AI writing has been found to occasionally contain inaccuracies in facts or references, while human writers tend to exercise greater caution in ensuring the accuracy of information. While AI writing is typically free of grammatical errors, it often lacks the emotional nuances and personal reflections characteristic of human writing [34]. However, exercising caution and considering ethical considerations is imperative when utilizing AI assistance in writing. In essay writing, AI can enhance efficiency by reducing time spent on writing [35], [36].

Additionally, AI can assist in detecting potential plagiarism, thereby ensuring the integrity and originality of scientific work [36]. AI can aid in grammar, style, and readability, ensuring clarity in communication [35]. Furthermore, AI can support statistical analysis and summarize complex information; however, the responsibility for the validity of the conclusions remains with the author [35]. It is imperative to acknowledge that AI cannot be recognized as an author, thereby necessitating the retention of full responsibility for the content by human authors [37].

Table 3. Rubrics of students' argumentation skills refers to Toulmin's Argumentation Pattern

Argumentation Skill	Assessment Scale			
	1	2	3	4
Claim	The claim is either unclear or irrelevant to the topic.	There is a claim but it does not focus on what should be the central claim.	The central claim is straightforward but can be interpreted in multiple ways or entirely irrelevant to the topic.	The central claim is obvious, can only be interpreted in one way, and is relevant to the topic being discussed
Data	There is no evidence, or the evidence is irrelevant to the claim.	There is evidence but the source is not entirely reliable, or it does not strongly support the claim.	The evidence comes from reliable sources and supports the claim, although the explanation could be improved.	The evidence comes from reliable sources and fully supports the claim with a strong explanation.
Warrant	There is no explanation, or the explanation does not connect the evidence and the claim logically.	The explanation does not fully explain how it relates to the claim.	The explanation shows the link between the evidence and the claim but it could be more detailed.	The explanation is obvious and logical. It links the evidence to the claim.
Backing	There is no additional explanation to strengthen the reasoning.	There is an additional explanation but it is weak.	The additional explanation supports the reasoning logically, although it could be more detailed.	The additional explanation is strong, relevant, and supports the reasoning.
Rebuttal	Fails to acknowledge or explain the weakness of the argument.	Acknowledges the weakness but the explanation is unclear or shallow.	Acknowledges and explains the weakness clearly but does not provide a detailed solution or improvement.	Acknowledges, explains the weakness, and offers a good solution or improvement.
Qualifier	There are no conditions or limitations for the claim.	There are conditions or limitations but they are unclear or vague.	Some conditions or limitations are explained clearly but not wholly.	Some conditions or limitations are apparent, specific, and detailed.

The ICC value of 0.621 indicates moderate to high reliability between raters, per the reliability criteria established by [28]. However, slight variations remain. The 95% confidence interval (0.429 - 0.777) indicates that the inter-rater reliability in the population ranges from moderate to high, with the lower limit falling within the moderate category. The *F*-Test results indicated a statistically significant inter-rater reliability ($p = 2.9\times10^{-9}$), substantiating that the observed consistency between peer raters was not merely a chance occurrence. The substantial reliability observed signifies that Toulmin's AS rubric can be employed to assess argumentative essays satisfactorily among raters. The comprehensive inter-rater reliability measurement outcomes can be found in Table 4.

Table 4. Inter-rater reliability measurement results

Parameter	Result
Model ICC	Two-Way Random Effects
Type	Absolute Agreement
Subjects (*n*)	31
Raters (*k*)	3
ICC (Single Measures)	0.621
95% CI	0.429 – 0.777
F-Test	$F(30, 61.3) = 5.83$
p-value	$2{,}9\times10^{-9}$

Students revised their essays based on self-assessment, peer feedback, collaborative discussions, and instructor input, producing final drafts evaluated by instructors. The instructor's assessment, the self-assessment, and the peer assessment were tested for Spearman correlation to assess the consistency of the assessment as learning of students' argumentation skills. The Spearman analysis result can be seen in Table 5.

Table 5. Spearman correlation analysis result

Variable	Spearman's rho (r_s)	*p*-value	Interpretation
Self-Assessment vs Peer Assessment	0,45799	0,00957	Moderate positive, significant.
Self-Assessment vs Instructor Assessment	0,42531	0,01707	Positive, significant
Peer Assessment vs Instructor Assessment	0,1367	0,46339	Positive, not significant

The correlation coefficient value, which ranges from 0.40 to 0.599, indicates that there is a moderate positive correlation between the way students rate themselves (self-rating) and how their friends rate them (peer-rating).

This relationship implies that students who rate themselves highly have a fairly high rating from their peers.

The p-value of 0.00957 shows that the observed relationship is statistically significant and unlikely to have occurred by chance. It suggests a correlation between how individuals perceive their abilities and how others perceive them. Previous research has shown that self-evaluations are often positively related to peer evaluations, although self-evaluations tend to be biased [38].

This moderate correlation also supports the view that students possess the capacity to evaluate the quality of their work. However, their self-assessment is not entirely consistent with the assessment of others, so further training in conducting self-assessment is needed to improve students' accuracy in assessing themselves.

The correlation analysis finding demonstrated a positive and significant relationship between student self-assessment and instructor assessment ($r_s = 0.42531$, $p = 0.01707$). This indicates that students who provide high self-assessment tend to provide higher-quality essays, reflecting the potential accuracy of self-assessment in assessing their abilities. This finding suggests that self-assessment can be a useful reflective tool to help students understand where to improve.

The correlation between peer and instructor assessment was weak ($r_s = 0.1367$) and not statistically significant ($p = 0.46339$). The preliminary and final essays showed no significant difference in scores. This finding supports the conclusion that students tend not to use peer assessment feedback effectively to improve essay quality. This result may be due to differences in peers' understanding of the essay evaluation standards or to students' tendency to give a more subjective score based on their perception of the essay without a direct link to the objective quality of the writing. These results also suggest that peer evaluation reflects students' perceptions of the first draft rather than their ability to revise based on feedback. The feedback may not have been critical or specific enough to be relevant to improving the quality of the essay. Based on the weak correlation results ($r_s = 0.1367$, $p = 0.46339$), peer assessment is more useful as a means of reflection for the feedback giver (essay rater). This finding suggests that the feedback from peer assessment is not critical or specific enough to improve the quality of writing, which affects its effectiveness as an assessment tool. Feedback must be more criterion-based and constructive to use peer assessment more effectively.

As demonstrated in Table 5, the outcomes of the present study suggest that employing self-assessment as a reflective tool could facilitate students' comprehension of their writing strengths and weaknesses.

However, its precision remains to be enhanced through further training. Conversely, peer assessment is more beneficial as a tool for introspection for the assessor than as a means of evaluating essay quality. Peer assessment feedback must be more targeted, constructive, and criterion-based to enhance its efficacy.

3.3. Argumentation Quality

Transcripts of collaborative argumentation on Padlet were analyzed using Toulmin's framework, with argument quality classified using a five-level model adapted from [29]. Examples of students' statements during collaborative argumentation discussions at each level are shown in Table 6.

Table 6. Examples of students' statements during collaborative argumentation discussions at each level

Level	Description	Example of Student Statement
1	Argument consists only of a simple claim (statement, conclusion, or opinion).	In my opinion, I agree with the main idea put forward by Dini. (Claim)
2	Argument includes a claim and at least one supporting reason or data.	Octobot has some disadvantages (Claim), for example, it has a very high cost and limited availability (Data).
3	Argument includes a claim, data, and a warrant (explains how the data supports the claim) or additional evidence/reasons (backing).	Pressure is one of the factors that affect the reaction rate (Claim). If pressure is increased, the volume of a gas decreases, resulting in more particles in the same space (Data). This leads to more frequent collisions between particles (Warrant), so the reaction rate increases (Backing).
4	Argument includes a claim, data, warrant or backing, and a qualifier (recognizes limits or exceptions).	Pressure has a significant effect on the reaction rate in the octobot system (Claim). This can be observed in the hydrogen peroxide decomposition reaction, which produces oxygen gas (Data). When pressure increases, the reaction proceeds more quickly (Warrant). However, if the pressure is too high, the catalyst may become uncontrollable, and if it's too low, the reaction may be too slow (Qualifier).
5	Argument includes a claim, data, warrant, backing, qualifier, and a rebuttal or consideration of an alternative view.	I disagree with the claim about the catalyst (Claim). The essay suggests that a stronger catalyst should be used, but I think that's not the best choice because the data doesn't consider the risk of losing control over the reaction (Rebuttal). Platinum, for instance, has medium efficiency (Data). If the catalyst is too strong, the reaction rate could be too fast, and if it's too weak, it could be too slow (Warrant). A medium-efficiency catalyst like platinum helps keep the reaction rate balanced (Backing). So, using platinum is better, as it avoids the risks of being either too fast or too slow (Qualifier).

Results showed the highest quality (Level 5) in 19.35% of cases, while most arguments reached Level 3 (29.03%), characterized by claims, data, and warrants. This finding indicates room for improvement in fostering high-quality arguments. The student's argument quality is shown in Figure 2.

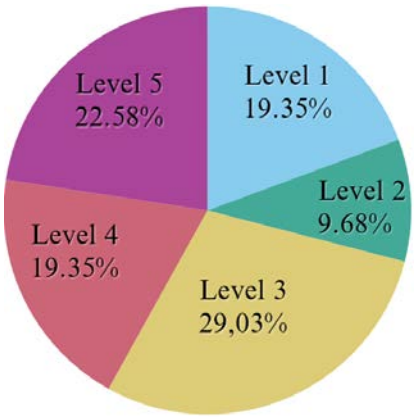

Figure 2. Students' argument quality

Integrating Toulmin's framework with collaborative tools like Padlet facilitated self-reflection, peer feedback, and collaborative discussions. These activities promoted logical argument construction, as evidenced by most students achieving level 3 argumentation skills.

The relatively low percentage of Level 5 arguments (19.35%) suggests that constructing high-level rebuttals remains challenging. This finding aligns with previous research indicating that rebuttals represent the most complex aspect of argumentation [29]. While the asynchronous format of Padlet allows time for reflection, it may not inherently support the dynamic exchange necessary for multiple rebuttals. Future interventions could focus on scaffolding strategies to help students anticipate and directly challenge the reasoning of opposing claims within a single, well-developed statement.

3.4. Practicality and Effectiveness

Questionnaires assessed the CASAL instrument's practicality and effectiveness. Practicality results (Figure 3) indicated moderate ratings for ease of use, clarity of instructions, and platform accessibility, with good ratings for completion time.

As shown in Table 7, open-ended responses highlighted areas for improvement, including simplifying instructions, extending task duration, and improving AI detection accuracy.

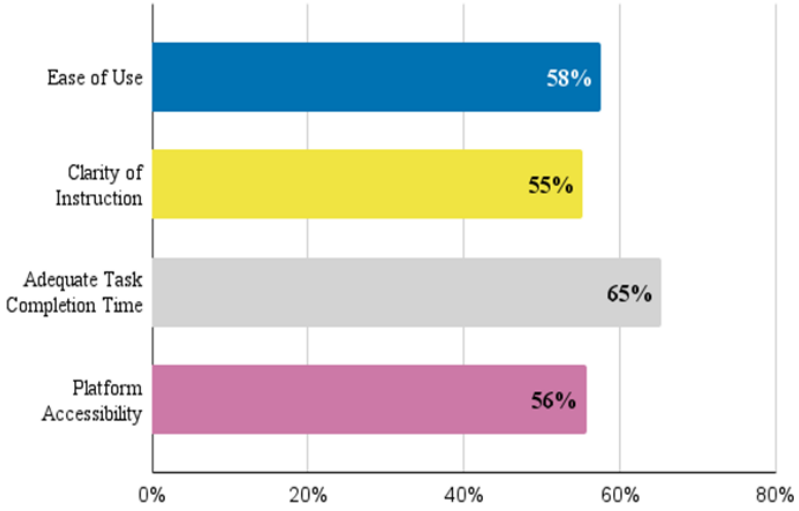

Figure 3. The CASAL instrument practicality

Table 7. Feedback on CASAL instrument through open-ended questionnaire

Aspect	Theme	Percentage	Description
The best part of the CASAL instrument	Self-Assessment	13%	It helps to realize weaknesses and strengths in AS and reflection for self-development.
	Peer Assessment	29%	It provides constructive criticism and suggestions.
	Media and platform	13%	Padlet supports discussion and collaboration.
	Motivation	19%	Increased student motivation through relevant and interesting topics.
	No Comments	26%	No comments or specific feedback.
The part needs to be Improved	Instructions and Rubrics	42%	Instructions are not clear enough and needs to be simplified.
	Implementation time	10%	Needs extended time
	AI detection	6%	Criticism of the accuracy of AI detection on essays
	No Comments	42%	Respondents felt there were no areas for improvement or no comments.
Other Suggestions	Task completion time	6%	Extended completion time to improve essay quality.
	Instrument and Rubric Improvement	39%	Make the instrument more structured and simplification of instruction.
	No Comments	55%	Respondents had no additional suggestions

Respondents appreciated the best parts of the CASAL instrument: self-assessment, peer assessment, media and platforms, and interesting topics. Self-assessment can help respondents identify weaknesses and strengths in AS and provide helpful reflections for self-development. These findings align with research [39] that self-assessment can improve metacognition, independent learning, and self-reflection so that students can recognize their strengths and weaknesses, motivate learning, improve time management, and promote ownership of learning, which ultimately supports continuous learning.

Meanwhile, peer assessment is beneficial in understanding the shortcomings of respondents' assignments through peer criticism and suggestions. Receiving assessment and feedback from friends is easier to understand than from teachers [40], so constructive suggestions can improve learning outcomes.

The phenomenon raised in this study, the factors affecting the rate of hydrogen peroxide decomposition reaction on Octobot, was appreciated by 19% of the students because it was considered relevant and interesting.

Previous research mentions that relevant and interesting topics can significantly increase learning motivation [41]. The best motivation for high learning outcomes is interest in the cognitive process and skill acquisition to promote persistent behavioural strategies and increase student engagement [41], [42].

Respondents also valued interactive media and platforms such as Padlet for supporting group discussions and facilitating collaboration.

Padlet is an AI-enabled online whiteboard that allows users to design the board's look more attractive and receive and give feedback to each other, making it easier for students and lecturers to have collaborative discussions. Padlet's use has been met with a favourable response from students [43], and assignments supported by the tool have been interesting [44].

Respondents provided feedback on three aspects of the system: the clarity of the instructions and rubric, the duration of the task, and the accuracy of the AI detector. The majority (42%) highlighted that the instructions on the rubric required simplification to facilitate understanding. In addition, respondents expressed concerns regarding the accuracy of the AI detection, as they perceived that, manual tasks were erroneously identified as AI results.

One key observation is the limited time allocated for essay revisions, which may have hindered students' ability to utilize feedback effectively.

This finding is consistent with the weak correlation between peer and instructor assessments, suggesting that peer feedback lacked the critical specificity needed for meaningful revisions. Further training in providing constructive, criteria-based feedback could enhance the utility of peer assessment. Additionally, the moderate correlation between self-assessment and other assessments indicates that students have some ability to evaluate their work but require additional training to improve accuracy and reduce bias.

As demonstrated in Figure 4, the data indicates that the effectiveness of the CASAL instrument is classified within the 'Moderate' category. 58% of respondents expressed that the CASAL instrument motivated them to improve their AS, while 54% moderately agreed that self-assessment supported self-reflection. Furthermore, 56% of respondents stated that peer feedback in peer assessment contributed to improving AS.

These outcomes align with the observations on students' argumentation skills in collaborative argumentation discussion activities on the Padlet platform, as presented in Figure 2. Most students' argumentation skills were classified as level 3, categorized as Fair. These results suggest that the CASAL instrument is moderately effective in training AS, though its performance could be enhanced.

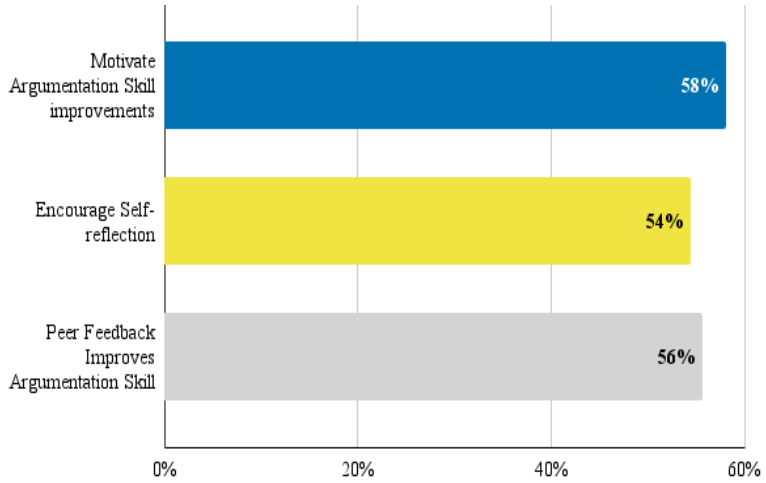

Figure 4. The CASAL instrument effectiveness

Analyzing students' initial and final essay drafts revealed no significant differences. This finding suggests that the time allocated between providing feedback and uploading the final essay draft was inadequate, potentially due to insufficient time for students to revise their essays. This finding aligns with the results presented in Table 7, which utilize a questionnaire to assess students' perceptions.

Another potential explanation could be that the feedback received from peers or instructors was unclear, irrelevant, or challenging to comprehend, hindering students' ability to revise effectively.

In order to ascertain the underlying cause of this phenomenon, it is needed to conduct a reflective interview session in which students can explain the reasons behind the decision to re-upload the initial draft.

Students' answers may include perceptions of the feedback received, obstacles to revising the essay, or even time constraints.

The development of the Instrument for Assessment as Learning Argumentation Skills (CASAL) demonstrated feasibility, exhibiting a highly valid category. However, in terms of practicality and effectiveness, it was still categorized as Fair.

Integrating Toulmin's framework with collaborative online whiteboard activities has facilitated self-reflection, peer feedback, and collaborative discussion, thereby contributing to students' ability to construct logical arguments. Nevertheless, the findings indicate the necessity for additional refinement in the feedback mechanism and training in self-assessment practices to enhance its practicality and effectiveness.

3.5. The Role Assessment as Learning in Improving Argumentation Skill

The CASAL instrument, developed within an assessment as learning approach, holds immense potential to promote students' argumentation skills through the integration of self-assessment, peer assessment, and collaborative argumentation. It allows students to take responsibility for learning and reflect on the arguments' quality. Indeed, the results showed that through self-assessment, students could indicate strengths and weaknesses in their essays, and 54% showed that their essays became more explicit and structured after reflection. It allows students to take ownership of their performance and learning outcome and identify what they must work on their learning [45]. Self-assessment and instructor assessment were moderately positively correlated ($r_s = 0.42531$, $p = 0.01707$); further guidance is needed for better self-assessment accuracy.

Integration of collaborative tools, such as Padlet, further reinforces the AaL approach by encouraging discussion and enhancing the development of logical arguments. Technology-supported collaboration, interaction, and knowledge-sharing positively influence metacognitive awareness and communication skills [46]. Peer assessment allows for collaborative learning since students gain insight into argumentation criteria when assessing the work of their peers. Furthermore, collaborative peer assessment has been found to enhance learning [47]. According to Toulmin's framework, peer interaction through collaborative peer assessment refines their claims and warrants.

However, the weak correlation between peer feedback and instructor evaluation at $r_s = 0.1367$, $p = 0.46339$ shows that more training is necessary to increase the quality and usefulness of peer feedback.

4. Conclusion

The development and implementation of the Chemistry Argumentation Skills Assessment for Learning (CASAL) instrument have demonstrated moderate practicality and effectiveness in improving the argumentation skills of pre-service chemistry teachers.

Integrating Toulmin's framework with collaborative online whiteboard activities facilitated self-reflection, peer feedback, and collaborative discussions, contributing to students' ability to construct logical arguments. However, the findings also highlight the need for further refinement in feedback mechanisms and increased training in self-assessment practices to enhance effectiveness.

The findings emphasize the potential for integrating chemistry education with self-assessment and collaborative argumentation activities, aligning with modern educational goals prioritizing 21st-century competencies. Digital platforms like Padlet show the role of technology in interactive learning experiences. The research provides valuable insights for curriculum developers and educators who want to use argumentation-focused assessments in their teaching.

However, there are limitations to the study that should be considered. First, it may not be representative of the broader population as the sample size was limited to 31 participants. Second, the short duration of the implementation stage may have limited student's ability to revise their essays effectively. Third, the quality of peer feedback varied, indicating a need for more rigorous training in delivering constructive and specific critiques. Future research should address these through more extensive and diverse samples, longer implementation times, and improved training modules on assessing peers.

Acknowledgements

This research was funded by the Ministry of Education, Culture, Research, and Technology of Indonesia under Project Title: "Pengembangan Instrumen Assessment as Learning untuk Meningkatkan Kemampuan Argumentasi Mahasiswa Calon Guru Kimia Menggunakan Collaborative Argumentation Online Whiteboard AI Assisted" under Contract Number: 071/E5/PG.02.00.PL/2024.

References:

[1]. Bisquert, P., Dupin de Saint-Cyr, F., & Besnard, P. (2019). Assessing arguments with schemes and fallacies. *International Conference on Logic Programming and Nonmonotonic Reasoning*, 61-74. Springer International Publishing. Doi: 10.1007/978-3-030-20528-7_6.

[2]. Giri, V., & Paily, M. U. (2020). Effect of scientific argumentation on the development of critical thinking. *Science & Education*, *29*(3), 673-690. Doi: 10.1007/s11191-020-00120-y.

[3]. Darmawansah, D., Lin, C. J., & Hwang, G. J. (2022). Empowering the collective reflection-based argumentation mapping strategy to enhance students' argumentative speaking. *Computers & Education*, *184*, 104516.

[4]. Zheng, X. L., et al. (2023). Effects of online whiteboard-based collaborative argumentation scaffolds on group-level cognitive regulations, written argument skills and regulation patterns. *Computers & education*, *207*, 104920. Doi: 10.1016/j.compedu.2023.104920.

[5]. Toulmin, S. E. (2003). *The Uses of Argument*. Cambridge University Press.

[6]. OECD. (2019). An OECD learning framework 2030. *The future of education and labor*, 23-35. Springer International Publishing. Doi: 10.1007/978-3-030-26068-2_3.

[7]. Burkard, A., et al. (2021). Argumentative skills: A systematic framework for teaching and learning. *Journal of Didactics of Philosophy*, *5*(2), 72-100. Doi: 10.46586/JDPh.2021.9599.

[8]. Suyono, S., et al. (2023). Student Support Program in Formulating a Research Framework. *The 10th Annual International Seminar on Trends in Science and Science Education (AISTSSE) FMIPA Universitas Negeri Medan.* Doi: 10.2478/9788367405782-033.

[9]. Martin, P. P., et al. (2024). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*, *61*(8), 1757-1792. Doi: 10.1002/tea.21903.

[10]. Erduran, S. (Ed.). (2019). *Argumentation in chemistry education: Research, policy and practice*. Royal Society of Chemistry.

[11]. Martins, M., & Macagno, F. (2022). An analytical instrument for coding and assessing argumentative dialogues in science teaching contexts. *Science Education*, *106*(3), 573-609. Doi: 10.1002/sce.21708.

[12]. Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. *Handbook of self-regulation*, 451-502. Academic Press.

[13]. Al-Ajmi, B., & Ambusaidi, A. (2022). The Level of Scientific Argumentation Skills in Chemistry Subject among Grade 11th Students: The Role of Logical Thinking. *Science Education International*, *33*(1), 66-74. Doi: 10.33828/sei.v33.i1.7.

[14]. Land, S., & Jonassen, D. (2012). *Theoretical Foundations of Learning Environments* (2nd ed.). Routledge. Doi: 10.4324/9780203813799.

[15]. Capkinoglu, E., Cetin, P. S., & Metin Peten, D. (2021). How do pre-service science teachers evaluate the persuasiveness of a socioscientific argument?. *International Journal of Science Education*, *43*(4), 594-623. Doi: 10.1080/09500693.2021.1876273.

[16]. Siswanto, S., et al. (2022). Infusing Explicit Argumentation in Science Reading Activities: Helping Prospective Science Teachers Reduce Misconception and Foster Argumentation Skills. *Pegem Journal of Education and Instruction*, *12*(3), 177-189. Doi: 10.47750/pegegog.12.03.19.

[17]. Rahayu, D. P., Wiyarsi, A., & Utami, A. (2022). Eksplorasi keterampilan argumentasi ilmiah mahasiswa melalui media sosial: topik kasus bioteknologi. *Jurnal Pendidikan MIPA*, *12*(3), 859-867. Doi: 10.37630/jpm.v12i3.703.

[18]. Gazali, F., et al. (2024). Profile of The First Year Student's Argumentation Skills on General Chemistry Courses at a Public University in West Sumatera: A Preliminary Study. *E3S Web of Conferences*, *481*, 04004. Doi: 10.1051/e3sconf/202448104004.

[19]. Yin, S., Chen, F., & Chang, H. (2022). Assessment as learning: How does peer assessment function in students' learning?. *Frontiers in Psychology*, *13*, 912568. Doi: 10.3389/fpsyg.2022.912568.

[20]. Dahal, N. D., et al. (2022). Enhancing student-teachers assessment skills: A self-and peer-assessment tool in higher education. *International Journal of Education and Practice*, *10*(4), 313-321. Doi: 10.18488/61.v10i4.31.

[21]. Han, J., et al. (2021). Learning analytics dashboards for adaptive support in face-to-face collaborative argumentation. *Computers & Education*, *163*, 104041. Doi: 10.1016/j.compedu.2020.104041.

[22]. Yan, Z. (2021). Assessment-as-learning in classrooms: the challenges and professional development. *Journal of Education for Teaching*, *47*(2), 293-295. Doi: 10.1080/02607476.2021.1885972.

[23]. Daniar, A. V., Herdyastuti, N., & Lutfi, A. (2023). Analysis effectiveness of implementation assessment as learning on metacognitive skills. *IJORER: International Journal of Recent Educational Research*, *4*(6), 759-770. Doi: 10.46245/ijorer.v4i6.392.

[24]. Rusmini, R., Suyono, S., & Agustini, R. (2021). Development of Critical Thinking Skills and Argumentation Skills Assessment Instruments Towards 21st-Century Skills Based on Non-Routine Problems. *International Joint Conference on Science and Engineering 2021 (IJCSE 2021)*, 382-389. Atlantis Press. Doi: 10.2991/aer.k.211215.066.

[25]. Yonata, B., Suyono, S., & 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. EDP Sciences. Doi: 10.1051/shsconf/202214901007.

[26]. Branch, R. M. (2009). *Instructional Design: The ADDIE Approach* (Vol. 772). New York: Springer.

[27]. Yusoff, M. S. B. (2019). ABC of content validation and content validity index calculation. *Education in medicine journal*, *11*(2), 49-54. Doi: 10.21315/eimj2019.11.2.6.

[28]. Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.

[29]. Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science education*, *88*(6), 915-933.

[30]. Plomp, T., & Nieveen, N. M. (2010). *An introduction to educational design research: Proceedings of the seminar conducted at the East China Normal University, Shanghai (PR China), November 23-26, 2007*. Stichting Leerplan Ontwikkeling (SLO).

[31]. Kar, S. K., et al. (2025). How sensitive are the free AI-detector tools in detecting AI-generated texts? A comparison of popular AI-detector tools. *Indian Journal of Psychological Medicine*, *47*(3), 275-278. Doi: 10.1177/02537176241247934

[32]. Akinwande, M., Adeliyi, O., & Yussuph, T. (2024). Decoding AI and Human Authorship: Nuances Revealed Through NLP and Statistical Analysis. *International Journal on Cybernetics & Informatics*, *13*(4), 85-103. Doi: 10.5121/ijci.2024.130408.

[33]. Revell, T., et al. (2024). ChatGPT versus human essayists: an exploration of the impact of artificial intelligence for authorship and academic integrity in the humanities. *International Journal for Educational Integrity*, *20*(1), 18. Doi: 10.1007/s40979-024-00161-8.

[34]. Georgiou, G. P. (2024). Differentiating between human-written and AI-generated texts using linguistic features automatically extracted from an online computational tool. *arXiv preprint arXiv:2407.03646*. Doi: 10.48550/arXiv.2407.03646

[35]. Tulandi, T. (2023). Disclosure of artificial intelligence/ChatGPT-generated manuscripts. *Journal of obstetrics and gynaecology Canada: JOGC= Journal d'obstetrique et gynecologie du Canada: JOGC*, *45*(8), 543-544. Doi: 10.1016/j.jogc.2023.03.013.

[36]. Scerbo, M. W. (2023). Can Artificial Intelligence Be My Coauthor?. *Simulation in Healthcare*, *18*(4), 215-218. Doi: 10.1097/SIH.0000000000000743.

[37]. Tomlinson, B., Torrance, A. W., & Black, R. W. (2023). ChatGPT and works scholarly: Best practices and legal pitfalls in writing with AI. *SMU l. Rev. f.*, *76*, 108. Doi: 10.48550/arXiv.2305.03722.

[38]. Panadero, E., & Brown, G. T. (2017). Teachers' reasons for using peer assessment: Positive experience predicts use. *European journal of psychology of education*, *32*(1), 133-156. Doi: 10.1007/s10212-015-0282-5.

[39]. Kyung, R. K. (2023). Going Beyond L2 Learning ─ The Effects of Self-Assessment on Student Metacognition, Self-Regulated Learning and Self-Reflection. *Korean Journal of General Education*, *17*(3), 201-206. Doi: 10.46392/kjge.2023.17.3.201.

[40]. Topping, K. J. (2023). Peer and Self-Assessment. *International encyclopedia of education*, 79-88. Elsevier.

[41]. Kolesnichenko, N. Y., et al. (2020). Increasing Students' Motivation to Learn at Tertiary Educational Institutions. *International journal of higher education*, *9*(7), 166-175.

[42]. Singh, M., et al. (2022). Impact of cognitive-behavioral motivation on student engagement. *Heliyon*, *8*(7). Doi: 10.1016/j.heliyon.2022.e09843.

[43]. Jong, B., & Tan, K. H. (2021). Using Padlet as a Technological Tool for Assessment of Students' Writing Skills in Online Classroom Settings. *International Journal of Education and Practice*, *9*(2), 411-423.

[44]. Gill-Simmen, L. (2021). Using Padlet in instructional design to promote cognitive engagement: a case study of undergraduate marketing students. *Journal of Learning Development in Higher Education*, (20).

[45]. Tsintsadze, T., Mskhaladze, D., & Dilaverova, T. (2024). The Benefits of Self–Assessment. *Proceedings of The International Education Conference*, *2*(1). Doi: 10.33422/ieconf.v2i1.620

[46]. Carvalho, A. R., & Santos, C. (2022). Developing peer mentors' collaborative and metacognitive skills with a technology-enhanced peer learning program. *Computers and Education Open*, *3*, 100070. Doi: 10.1016/j.caeo.2021.100070.

[47]. Rico-Juan, J. R., Cachero, C., & Macià, H. (2022). Influence of individual versus collaborative peer assessment on score accuracy and learning outcomes in higher education: an empirical study. *Assessment & Evaluation in Higher Education*, *47*(4), 570-587. Doi: 10.1080/02602938.2021.1955090.