



From Preparation to Public Response: Think-Pair-Share and Hot Seat Sequences to Support Argumentative Speaking Skills in Grade XI Students

Fanisa Maulidiyah^{1*}, Muhammadu Sainulabdeen Zunoomy², & Muhammad Farris Imadi Fuze³

¹ Independent Researcher, Pasuruan, Indonesia

²South Eastern University of Sri Lanka, Oluvil, Sri Lanka

³Chungnam National University, South Korea

Corresponding author: fanisamaulidiyah@gmail.com

Article History

Submitted: 07-09-2026

Revised: 23-09-2026

Accepted: 24-09-2026

Published: 25-09-2026

Keywords:

argumentative speaking;
Think-Pair-Share; Hot Seat;
classroom action research;
accountable talk

Abstract

Argumentative speaking requires students not only to produce linguistically appropriate utterances but also to take a stance, use evidence, construct reasoning, and respond to questions. This classroom action research examined changes in argumentative speaking performance and learning process indicators through an integrated Think-Pair-Share (TPS) and Hot Seat sequence. The study involved 33 Grade XI students in an Indonesian language class at SMA Islam Pasuruan during the second semester of the 2025/2026 academic year. Data were collected through speaking performance assessments, structured observations, learning process rubrics, questionnaires, semi-structured interviews, field notes, and documentation. Speaking performance was assessed using nine analytic indicators, while learning processes were examined through five indicators. The mean speaking performance score increased from 70.13 in Cycle I to 77.70 in Cycle II, with 30 students (90.9%) achieving ≥ 75 . The largest improvements occurred in Gestures and Expressions (+0.79) and Content Mastery (+0.58), whereas Diction and Grammar remained unchanged (2.91). The mean learning process score increased from 62.42 to 77.88, with notable gains in Deep Learning Achievement (+0.70), Self-Efficacy (+0.67), and Active Interaction (+0.63). Cycle II incorporated a shared topic, Claim–Evidence–Conclusion structure, mandatory oral rehearsal, extended Pair time, Indonesian-only interaction, targeted scaffolding, and audience Q&A. The findings suggest that sequencing peer preparation and public response may support argumentative and public speaking performance, although linguistic accuracy did not improve correspondingly.



© 2026 by the authors. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International License (CC BY 4.0).

1. INTRODUCTION

Speaking argumentatively requires more than merely producing a grammatically correct speech. Students must take a stance, select relevant evidence, link that evidence to their claim, construct a line of reasoning, and respond to audience questions unscripted. The Toulmin model highlights the relationship between claims, data, and warrants, while studies on speaking assessment emphasize fluency, comprehension, vocabulary, grammar, and pronunciation as dimensions of oral performance (Toulmin, 2003; Brown, 2004). These demands make argumentative speaking an ideal lens through which to observe how classroom interaction can support both cognitive preparation and immediate performance in the public space of a classroom.

However, opportunities to speak do not automatically result in high-quality argumentation. The quality of classroom interaction, including opportunities to explain, ask questions, respond, and revise one's own thinking, lies at the heart of language and cognitive development (Hall, 2022; Sybing, 2021). Recent studies on academically productive classroom discourse confirm that dialogue meeting standards of accountability to the learning community, to norms of reasoning, and to knowledge yields better learning outcomes than lively but unstructured talk (Hennessy et al., 2023). Research on foreign

language learning also indicates that speaking anxiety can affect both performance and the learning experience (Horwitz et al., 1986; Zhang, 2019; Veenstra & Weaver, 2022), a pattern that, according to a recent review of foreign language anxiety research, remains relevant across learning contexts (Yu, 2024). These findings explain why students require preparation before facing more demanding speaking tasks and why instruction needs to balance cognitive preparation with oral interaction that increases in intensity over time.

Think-Pair-Share offers a useful preparation structure: students formulate ideas individually, test them with a partner, and share them with a wider audience. However, the "Share" phase does not inherently guarantee equitable or sufficiently challenging participation in the discussion. Previous research has identified disparities in the distribution of speaking opportunities during the "Share" phase, noting a tendency for it to be dominated by a handful of confident students (Cooper et al., 2021). Other studies report that TPS can increase participation, promote a more equitable distribution of participation, and improve discussion quality (Guenther & Abbott, 2024), as well as oral communication (Raba, 2017); together with the concerns above, this suggests that outcomes depend on how the Share stage is structured. The "Hot Seat" offers a potential solution to this gap by creating a structured scenario in which one student must explain, clarify, and defend a position while classmates pose questions. It must be emphasized at the outset that the literature specifically examining the "Hot Seat" as a pedagogical technique remains limited, consisting largely of student theses or small-scale local journal articles that have not undergone rigorous peer review. Consequently, this study does not base its theoretical justification on "Hot Seat" literature itself, but rather on theories of dialogic discourse and "accountable talk" (Hennessy et al., 2023), whose empirical support comes partly from early childhood classrooms (van der Veen et al., 2021; van der Wilt et al., 2022); extending this support to Grade XI students is an assumption that this study does not test. Research on classroom debate indicates that the communicative goal of a debate shapes how students argue (González Palta et al., 2026), and chatbot-assisted in-class debates have been examined for their effects on argumentation skills and task motivation (Guo et al., 2023). Meta-analytic evidence regarding "productive peer talk" also indicates that structured discourse moves can enhance the quality of interaction and learning outcomes during peer-based activities (Hu & Chen, 2024).

This study originated from preliminary observations at an Islamic high school in Pasuruan, where students exhibited passivity, difficulty organizing ideas, reliance on written notes, and limited confidence when asked to respond spontaneously. This class was selected based on the aforementioned observations and consultations with the subject teacher, both of which indicated clear difficulties regarding speaking skills and engagement in discussions during the argumentative text unit. A pre-intervention assessment conducted on January 14, 2026, involving 33 students, yielded an average speaking score of 49.94, with not a single student reaching the mastery threshold of 75. This score served as a diagnostic snapshot of the initial conditions. As the pre-intervention rubric differed from the one used during the two intervention cycles, the baseline score was not used for a direct comparison of indicator levels against the cycle scores. This study addresses a specific gap in classroom-based evidence: rather than examining Think-Pair-Share (TPS) or Hot Seat as standalone activities, it investigates the two as a unified instructional sequence progressing from preparation to response. TPS is positioned as a space for cognitive and social preparation through the stage of "Think" and "Pair" serve as preparatory phases, whereas "Hot Seat" acts as a performative stage that tests readiness through argumentative responses delivered before an audience. In this article, the term "accountability bridge" is employed as an analytical concept derived from the study's findings, rather than as an established theoretical construct. The concept refers to the function of the "Hot Seat" in linking the outcomes of individual and peer preparation with the demands of listening, explaining, defending claims, and responding to questions in a public setting.

Based on this background, the following research questions were formulated:

- 1) How did the argumentative speaking performance of Grade XI students evolve from the pre-action stage through Cycle I to Cycle II?
- 2) How did the indicators of their learning experience processes change across the two action cycles?

Thus, the study's contribution is not a claim that TPS or the Hot Seat in isolation caused the improvement; rather, it offers classroom-based evidence regarding how these two activities can be structured to create a more accountable interactional progression.

2. THEORETICAL FRAMEWORK

2.1 Argumentative Speaking Skills

Argumentative speaking integrates propositional content, reasoning, linguistic resources, and interactional performance into a unified whole. In this study, the construct of speaking performance is operationalized through nine indicators analytically grouped into three domains: content and argumentation quality (Content Mastery, Argument Organization, Argument Structure and Content, and Responsiveness and Assertiveness); linguistic performance (Fluency, Diction and Grammar, and Pronunciation and Intonation); and public delivery (Self-confidence and Gestures and Expressions). These nine indicators draw upon several complementary frameworks: Brown's (2004, 2007) linguistic dimensions, Toulmin's (2003) argument structure, Padmawati et al.'s (2019) dimensions of speaking performance (including expressions and gestures), Bandura's (1977) construct of self-confidence, and Zarefsky's (2009) concept of assertiveness in responding to rebuttals. This grouping serves to clarify the function of each indicator rather than to claim that the three domains represent empirically validated psychometric factors within this study's sample.

As some indicators are conceptually close, the distinctions between them are defined by the focus of the assessment. Content Mastery evaluates the accuracy and depth of the content and data presented; Argument Organization assesses the coherence and interconnectedness of ideas during the delivery of arguments; Argument Structure and Content evaluates the clarity of claims as well as the adequacy and relevance of supporting evidence; and Responsiveness and Assertiveness assess the ability to respond to questions or rebuttals directly and confidently. Fluency, Diction and Grammar, and Pronunciation and Intonation focus on linguistic performance, whereas Self-confidence and Gestures and Expressions focus on public delivery. Conceptual overlap among the indicators remains a limitation of the rubric. Therefore, indicator scores are used primarily to observe patterns of change across cycles, rather than to infer factor structure or discriminant validity.

2.2 Think-Pair-Share as Preparation and Peer Scaffolding

TPS allows students to transition from generating individual ideas to peer testing and public sharing. The "Think" stage provides space for students to identify their positions and supporting reasons; the "Pair" stage offers an opportunity to test the coherence of that reasoning and receive peer feedback; and the "Share" stage extends the interaction to a wider audience. Research on cooperative learning indicates that structured peer interaction can sustain participation and learning when roles and accountability are made explicit (Gillies, 2016). A recent systematic review of "productive peer talk moves" further confirms that it is the quality not merely the presence of peer interaction that determines learning outcomes; thus, TPS needs to be designed with clear accountability mechanisms rather than being allowed to proceed as a mere formal routine (Hu & Chen, 2024). A recent study in medical education found that TPS increased participation and promoted a more equitable distribution of participation, with students identifying thinking and pairing time as especially important (Guenther & Abbott, 2024). This is consistent with the decision in Cycle II to make paired oral practice mandatory rather than leaving it voluntary. In this study, TPS is treated as a scaffold rather than a standalone speaking task.

2.3 The Hot Seat as Responsive Argumentation

The Hot Seat functions as a structured oral response activity in which a selected student presents a position and responds to questions or counterarguments from the audience. As noted in the Introduction, literature specifically addressing the "Hot Seat" as a teaching technique remains scarce and is dominated

by small-scale reports that may not meet the rigorous standards of reputable journals. Consequently, this study grounds the pedagogical value of the Hot Seat in broader theories of dialogic and accountable talk, rather than in the literature on the technique itself.

The "accountable talk" framework posits that academically productive classroom discourse must be accountable to the learning community, to norms of reasoning, and to knowledge itself; these three dimensions require students to listen, build upon the ideas of others, and defend claims with verifiable reasoning (Hennessy et al., 2023). The Hot Seat can be viewed as a concrete manifestation of these three forms of accountability: the student in the hot seat cannot simply recite memorized arguments but must listen carefully to audience questions, select relevant aspects of their argument, and defend their position spontaneously. Evidence from the broader fields of classroom debate and oral argumentation is consistent with this logic: the communicative goal of a debate shapes how students argue (González Palta et al., 2026), and chatbot-assisted in-class debates have been examined for their effects on argumentation skills and task motivation (Guo et al., 2023). Thus, the pedagogical value of the Hot Seat in this study lies in the heightened interactional demands that follow the opportunity to prepare ideas rather than in the pressure itself being inherently beneficial. This study views the Hot Seat as valuable precisely when such pressure is preceded by adequate preparation and peer practice, allowing the activity to serve as a bridge of accountability that transforms the outcomes of Think-Pair-Share (TPS) exercises into public argumentative performances subject to real-world scrutiny.

2.4 Learning Experience and Self-Efficacy

Here, learning experience is understood as an experience manifested through cognitive, affective, and social engagement during the learning process. Dewey (1997) emphasized the continuity between experience and learning, while Bandura (1977) explained how experiences of success and social feedback can shape beliefs regarding one's own capabilities. Research on EFL public speaking has also identified multiple sources of learners' self-efficacy, including prior experience, social persuasion, and contextual support (Zhang & Ardasheva, 2019). This study tracks five process indicators: Reorganization of Experience, Deep Learning Achievement, Critical Analysis, Self-Efficacy, and Active Interaction derived from distinct theoretical frameworks (respectively: Dewey, 1997; Fullan, 2013; Kolb & Kolb, 2017; Bandura, 1977; Vygotsky, 1978). The instrument was developed by the researchers, and its content validity was reviewed by Indonesian language education experts/practitioners and subject matter teachers. These five indicators are treated as a set of complementary process indicators rather than as evidence that "learning experience" constitutes a single psychometric construct. Specifically, the self-efficacy indicator is interpreted cautiously as one of the process indicators and is not used to make psychometric claims about self-efficacy as a standalone construct. Empirical evidence regarding public-speaking self-efficacy also indicates a relevant relationship between self-belief and speaking performance, although this relationship cannot be tested for causality within the current research design (Zhang et al., 2020).

2.5 Preparation-to-Response Sequence

Based on the synthesis of the framework above, this study views argumentative speaking instruction as a staged progression from preparation to public response. The "Think" stage provides time to construct a position and supporting reasons; the "Pair" stage allows students to test, refine, and rehearse arguments with peers; the "Hot Seat" then increases interactional demands, as students must respond directly to audience questions. Within this framework, the "Hot Seat" is not treated as a technique effective in isolation, but rather as a point for testing readiness built during the preceding stages. This sequence establishes a bridge of accountability: students move from accountability for their own ideas to accountability to a partner, and finally to accountability to an audience through responses subject to scrutiny. This framework serves as an interpretive lens for the research rather than a causally tested model.

3. METHOD

3.1 Design and Context

This study employed classroom action research (CAR), following the cyclical logic of planning, action, observation, and reflection (Kemmis et al., 2014). Two action cycles were conducted with 33 Grade XI students (8 male, 25 female) at SMA Islam Pasuruan during the second semester of the 2025/2026 academic year. The researcher acted directly as the Indonesian language teacher for the class (teacher-researcher), while another Indonesian language teacher from the same school served as an independent observer and collaborator, as well as a second rater for inter-rater reliability purposes (see Section 3.2). This class was selected based on preliminary observations and consultations with the subject teacher, which indicated difficulties regarding speaking skills and discussion participation during the argumentative text unit. The aim of this CAR was to improve classroom practice in response to the observed difficulties, rather than to estimate the population-level effects of an intervention.

The class was an Indonesian language (Bahasa Indonesia) class; Indonesian was the language of instruction and the target language of the speaking tasks, and many students switched to Javanese during pair discussions in Cycle I (Table 1). Literature on foreign-language anxiety and EFL public speaking is therefore used as an analogy for speaking in a language that is not always the students' preferred language in informal interaction, not as a claim that Indonesian is a foreign language for these students.

3.2 Data Sources and Instruments

Speaking performance was assessed using a nine-indicator analytic rubric adapted from Brown's speaking assessment dimensions and Toulmin's argument structure, with full descriptors for levels 1–4. The use of analytic rubric formats is also consistent with research emphasizing their value for assessing oral presentation performance in secondary education (Nadolski et al., 2021).

The pre-intervention phase utilized Brown's (2004) five-indicator rubric (using a 1–5 scale) covering grammar, vocabulary, fluency, pronunciation, and comprehension whereas assessments in Cycles I and II employed a nine-indicator analytic rubric (using a 1–4 scale) developed to address the specific demands of oral argumentation. Although results from both phases could be reported on a 0–100 scale, this shared reporting scale did not render the two instruments equivalent in terms of constructs, indicators, or scoring ranges. Consequently, pre-intervention scores served primarily to establish the class's diagnostic baseline, while the analysis of changes in indicator levels and overall performance focused more robustly on comparing Cycle I and Cycle II results using the same instrument.

Content validity was reviewed on January 8, 2026, by an Indonesian language education expert/practitioner, who also served as the thesis supervisor and the second author of this article, and by the Indonesian language teacher at the partner school. The second author's involvement as a validator is transparently disclosed. Validation efforts focused on the instrument's alignment with the research objectives, argumentative speaking competencies, and clarity of instructions. For the speaking assessments conducted during the intervention cycles, the researcher and subject teacher independently evaluated all 33 students using the same rubric. Discrepancies in scoring arose for 13 students regarding specific indicators, particularly gestures, assertiveness, and fluency. Both raters subsequently reviewed video recordings together to reconcile the differing scores; the reconciled scores were used in the final results. Paired raw scores prior to reconciliation were not stored in a format that allowed for the retrospective calculation of ICC or weighted kappa. Consequently, this study does not report numerical reliability coefficients and does not treat reconciliation as a substitute for evidence of inter-rater reliability.

3.3 Action Procedure

This intervention was conducted through two action cycles following the TPS and Hot Seat sequences. Each cycle comprised the stages of preparation, student–student interaction, public response, observation, and reflection. Cycle I was conducted to identify student responses to the initial procedures and uncover obstacles affecting their argumentative speaking performance and participation. Based on the observations and reflections from Cycle I, several procedural modifications were implemented in Cycle II, specifically regarding topic selection, argument structure, oral practice, language use, audience

participation, and instructional support (scaffolding) from the teachers. Table 1 outlines the characteristics of the procedures and modifications made in both cycles.

Table 1.

Procedural characteristics, observations, and modifications across the two action cycles.

Aspect	Cycle I (January 21, 2026)	Cycle II
Topic	Five different topics: (1) online vs. offline group learning, (2) helmet requirements when cycling/light sports, (3) parents as a place to confide, (4) the role of a college diploma in the world of work, (5) formal education vs. future skills	One common topic: “ <i>Digital World Dilemma: Is Social Media Making Us Smarter or Lazier?</i> ”
Argument structure	Not explicitly required	Required, built through three explicit components: Claim, Evidence, and Conclusion
Think Stage	15 minutes, individual, recording claims and supporting data	10 minutes
Pair Stage	10 minutes, discuss and agree on one argument together	15 minutes, including mandatory oral practice and a simulated Q&A session
Language of instruction	Many students switched to Javanese during pair discussions.	Indonesian is required throughout the activity
Hot Seat Selection	Randomly selected by the teacher	Randomly selected; three students who needed intensive support received scaffolding directly from the teacher
Audience participation	Not required; predominantly extroverted students	The audience is required to ask questions, and the teacher awards small appreciation points to encourage participation
Observational findings	Students still rely on notes, provide very short answers, make minimal eye contact, often code-switch to Javanese, and participate in Q&A sessions unevenly.	More thorough preparation, increased willingness to respond, and more active audience participation
Reflection and modification for the next cycle	Three main barriers were identified: first language interference, dominance of extroverted students, visual reliance on notes	The above modifications were designed based on these three findings

3.4 Data Analysis

Quantitative data are summarized using means, mastery percentages, and absolute score changes at baseline and between cycles. Changes at the indicator level were also examined to identify which dimensions showed stronger or weaker shifts. Rubric means on the 1–4 scale were converted to a 0–100 scale as $(\text{mean score} / 4) \times 100$, so a mean of 3.00 corresponds to the mastery threshold of 75. Learning experience scores are reported only for Cycle I and Cycle II, as no baseline measurement existed prior to the intervention. Qualitative data from field notes, interview findings, and documentation were analyzed interpretively by comparing findings across sources and discussing them with collaborators during the reflection stage of each cycle. This analysis serves to explain patterns observed in the quantitative data rather than to generate standalone thematic claims. As this study did not employ a formal thematic coding protocol involving multiple independent coders, qualitative evidence is treated as supplementary, and this limitation is explicitly stated.

3.5 Research Ethics

This study was conducted as part of routine classroom instruction with official research authorization from the relevant institution (No. 28.11.15/UN32.2.1/dk/2025). The intervention was implemented within the context of regular lessons and did not require students to participate in activities

outside standard classroom instruction. To protect participant privacy, all student identities were anonymized in the research records, analyses, and the article manuscript itself. Photographs and video recordings were taken solely to document classroom activities and support the verification of the research process. These materials are securely stored by the researcher, are not publicly disseminated, and are accessible only for legitimate research. No names, personally identifiable information, or other identifying details regarding individual students were disclosed in this study.

4. RESULTS

4.1 Baseline Conditions

The pre-action assessment provided a diagnostic snapshot of the class's condition prior to the intervention. The mean speaking score was 49.94, and none of the 33 students met the threshold of 75. The pre-action assessment utilized Brown's five-indicator rubric (grammar, vocabulary, fluency, pronunciation, and comprehension), whereas the assessments for Cycles I and II employed a nine-indicator rubric. Due to the difference in instruments, the pre-action results were not treated as fully equivalent to the cycle scores; this baseline figure served primarily to illustrate the severity of the problem that prompted the intervention.

Table 2.

Descriptive Overview of Speaking Performance at Pre-Action, Cycle I, and Cycle II

Measure	Pre-action	Cycle I	Cycle II
Average speaking score (class)	49.94	70.13	77.70
Students achieving ≥ 75	0/33 (0%)	not used as a final criterion	30/33 (90.9%)
Instruments used	Brown's 5-indicator rubric (scale 1–5)	Researcher-developed 9-indicator rubric (scale 1–4)	Researcher-developed 9-indicator rubric (scale 1–4)

Note: The pre-intervention score served as a diagnostic snapshot. Because the pre-intervention rubric differed from the rubrics used in Cycles I and II, the most methodologically sound comparison of changes in indicator levels is that between Cycles I and II.

Descriptive results for all three time points are shown in Table 2 and interpreted in Section 4.2.

4.2 Changes in Argumentative Speaking Performance

The mean speaking scores were 49.94 at pre-intervention, 70.13 in Cycle I, and 77.70 in Cycle II. Descriptively, this represents an increase of 20.19 points from pre-intervention to Cycle I and 7.57 points from Cycle I to Cycle II. However, the increase from pre-intervention to Cycle I must be interpreted with caution, given that the two stages employed different instruments for data collection. A more instrumentally robust comparison is the change between cycles, where the same nine-indicator rubric was applied consistently. As this was a single-class CAR study without a comparison group, these changes were interpreted as documented shifts throughout the sequence of actions rather than controlled causal estimates of effectiveness.

Table 3.

Changes in indicator levels regarding argumentative speaking performance

Indicator	Cycle I	Cycle II	Change
Content Mastery	2.88	3.46	+0.58
Argument Organization	2.61	2.97	+0.36
Fluency	2.70	2.91	+0.21
Diction and Grammar	2.91	2.91	0.00

Indicator	Cycle I	Cycle II	Change
Pronunciation and Intonation	2.88	3.00	+0.12
Self-confidence	2.88	3.10	+0.22
Gestures and Expressions	2.64	3.43	+0.79
Argument Structure and Content	2.82	3.16	+0.34
Responsiveness and Assertiveness	2.94	3.07	+0.13
Class average (scale 0–100)	70.13	77.70	+7.57

The largest inter-cycle improvement occurred in Gestures and Expressions (+0.79), followed by Content Mastery (+0.58). Argument Organization (+0.36) and Argument Structure and Content (+0.34) also improved in this study. Fluency (+0.21), Self-confidence (+0.22), Pronunciation and Intonation (+0.12), and Responsiveness and Assertiveness (+0.13) showed smaller gains. Diction and Grammar remained unchanged at 2.91. Consequently, the inter-cycle changes were uneven, with non-linguistic and content-related dimensions shifting more significantly than technical linguistic accuracy.

Qualitative evidence provides a concrete picture of this pattern. In Cycle I, field notes recorded that three students (Students A–C) gave very brief "Hot Seat" responses and relied heavily on notes, whereas three others (Students D–F) appeared more willing to ask questions and respond spontaneously without leaning on written notes. In Cycle II, the three students who previously required extra guidance (Students A–C) received direct scaffolding from the teacher and showed score increases; however, all three remained in the "Not Yet Competent" category (score of 50) at the end of Cycle II—a finding that underscores the need for sustained additional support for students with specific needs, rather than support limited to a single cycle.

Learning experience process indicators were measured at the end of each cycle using a five-indicator rubric. The class average rose from 62.42 in Cycle I to 77.88 in Cycle II, an increase of 15.46 points. Twenty-nine out of 33 students (87.9%) reached the threshold of 75 by the end of Cycle II. As there was no baseline for the learning experience, this change is interpreted solely as an inter-cycle shift. The five indicators stem from distinct theoretical foundations and are therefore interpreted as a set of complementary process indicators rather than as a single psychometric construct.

4.3 Changes in the Learning Experience

Deep Learning Achievement showed the greatest increase (+0.70), followed by self-efficacy (+0.67) and active interaction (+0.63). Experience reorganization and critical analysis each improved by more than half a scale point. Observational evidence provides context for these patterns: in Cycle I, students often required prompting to take a stance, relied on their notes, and used their first language during pair discussions. Following the revisions in Cycle II, pair exercises became mandatory, topics were narrowed to a single shared issue, the claim–evidence–conclusion structure was explicitly explained, and audience Q&A was required. Observers noted better preparation, a greater willingness to respond, and more active audience participation.

Table 4.

Indicator-level changes in learning experiences.

Indicator	Cycle I	Cycle II	Change
Experience Reorganization	2.79	3.33	+0.54
Deep Learning Achievement	2.33	3.03	+0.70
Critical Analysis	2.39	2.94	+0.55
Self-Efficacy	2.42	3.09	+0.67
Active Interaction	2.55	3.18	+0.63
Class average (scale 0–100)	62.42	77.88	+15.46

Table 4 shows that all five process indicators rose between cycles, but from different starting points. Deep Learning Achievement had the lowest mean in Cycle I (2.33) and gained the most, while Self-Efficacy (2.42), the third lowest, gained the second most, so part of these larger gains may reflect greater room for improvement. Experience Reorganization was the highest indicator in both cycles (2.79 and 3.33). By the end of Cycle II, four of the five indicators had a mean of at least 3.00, which corresponds to a score of 75 when means are expressed on a 0–100 scale ($\text{mean}/4 \times 100$). Critical Analysis (2.94) remained just below this level, suggesting it is the process dimension most in need of further support. This comparison is descriptive only, as there was no baseline measurement and differences between indicators were not tested.

4.4 Process Changes Between Cycles

Qualitative notes indicate that the most significant procedural shift lay not in the basic TPS structure, but in how accountability was built into it. In Cycle I, students could complete the pair interaction without practicing orally, and whole-class Q&A tended to be dominated by extroverted students. In Cycle II, the pair stage included mandatory oral practice, the audience was required to ask questions, and teacher scaffolding was specifically directed at students needing extra support. These changes coincided with stronger improvements in gestures, content mastery, self-efficacy, deep learning, and active interaction.

5. DISCUSSION

5.1 From Preparation to Public Response

These findings clarify the role of the "Hot Seat" within the learning sequence. The Hot Seat is not merely a turn to present in front of the class; rather, it is a stage where the quality of preparation is tested against questions that cannot be fully predicted. When the audience is required to ask questions, students must listen, select relevant parts of their arguments, and defend their positions in real time. This function supports the conceptualization of the Hot Seat as a bridge for accountability. However, the research data do not allow for isolating the Hot Seat's contribution from that of the Think-Pair-Share (TPS) activity, practice sessions, teacher feedback, and other changes in task conditions.

Cross-cycle patterns reveal performance changes following the implementation and subsequent revision of the learning sequence. The 7.57-point increase from Cycle I to Cycle II coincided with several targeted modifications: mandatory oral practice during the "Pair" stage, the use of a shared topic, a Claim–Evidence–Conclusion structure, and heightened audience accountability. Improvements in Argument Organization (+0.36) and Argument Structure and Content (+0.34) align with the function of the "Think" and "Pair" stages as spaces for constructing and testing arguments prior to public response. However, because these changes were introduced simultaneously, the findings cannot determine the specific contribution of each component.

Consequently, the research design better supports the pedagogical viability of a "preparation-to-response" sequence than claims that TPS or the Hot Seat independently drive improvement. Students experienced both activities as a unified learning sequence, while Cycle II also involved changes to task conditions and the nature of teacher support. The study's primary contribution lies in the principles of sequencing and building accountability, rather than in isolating the effects of individual techniques.

5.2 Gestures and Content Mastery Improved Most Sharply

Gestures and Expressions showed the largest increase in indicator levels (+0.79). This pattern

aligns with the nature of repeated public performance: students transitioning repeatedly from "Pair" exercises to the "Hot Seat" had more opportunities to manage eye contact, posture, facial expressions, and vocal delivery. Content Mastery also improved sharply (+0.58), consistent with Cycle II's requirement for students to construct arguments using Claims, Evidence, and Conclusions centered on a single shared topic. A focused topic reduced the cognitive load associated with switching between issues, allowing students to deepen their preparation.

However, this interpretation remains a process-level explanation rather than an experimental mechanism. The action cycle combined practice, teacher feedback, repeated exposure, and evolving task conditions, all of which could plausibly contribute to the observed improvements.

5.3 Diction and Grammar Did Not Change

Diction and Grammar remained unchanged at 2.91 across Cycle I and Cycle II, despite improvements in most other speaking indicators. This finding suggests that increased opportunities for argument construction, peer rehearsal, and public response may not automatically produce gains in linguistic accuracy. The Cycle II revisions emphasized preparation, argument structure, oral practice, and interactional responsiveness rather than sustained form-focused instruction on vocabulary and grammar. The unchanged score therefore highlights an important pedagogical distinction: students may become more confident and effective in organizing and delivering arguments without showing a corresponding improvement in grammatical and lexical accuracy. Explicit and sustained language-form instruction may be necessary when linguistic accuracy is a primary learning target.

5.4 Learning Experience as Process Evidence

The relationship between the learning experience process indicators and speaking performance must be interpreted with caution. Learning experience was not measured at baseline, the five indicators stemmed from different theoretical foundations, and no analysis was designed to test statistical relationships or mediation between the constructs. Consequently, this study concludes that both learning experience process indicators and speaking performance showed improvement throughout the cycles, not that changes in learning experience caused changes in speaking performance.

The increase in learning experience scores from 62.42 to 77.88 offers complementary evidence that process indicators shifted across the cycles. Deep Learning (+0.70) and Self-Efficacy (+0.67) showed the greatest gains. The "Pair" stage likely provided a relatively safer space for students to test ideas before performing publicly, while the requirement for the audience to ask questions created a demand for attentive listening to classmates' arguments. This interpretation is process-oriented and supported by classroom observations, rather than constituting evidence of a causal or mediational relationship between learning experience and speaking performance.

5.5 Cycle II Revisions as a Pedagogical Mechanism

The Cycle II modifications were not arbitrary additions; each responded to a barrier identified in the Cycle I reflection (Table 1). First-language interference during pair discussions was addressed by requiring Indonesian throughout the activity. The dominance of extroverted students in whole-class Q&A was addressed by making audience questions mandatory and awarding small appreciation points. Reliance on notes and very short Hot Seat answers was addressed through a required Claim–Evidence–Conclusion structure, a single shared topic, mandatory oral rehearsal with a simulated Q&A in the Pair stage, and direct teacher scaffolding for three students who needed additional support. Read this way, Cycle II strengthened accountability at each stage of the sequence: to one's own reasoning in Think, to a partner in Pair, and to an audience in the Hot Seat. Because these changes were introduced together, however, the design cannot show which of them contributed most. This account is a plausible process-

level explanation rather than a tested mechanism.

5.6 Contribution of the Study

In contrast to previous TPS research highlighting the unequal distribution of speaking opportunities during the "Share" phase (Cooper et al., 2021), this study contributes a concrete sequencing principle for argumentative speaking: cognitive preparation precedes peer testing, and peer testing precedes more demanding public responses. This sequence is not a novel method but rather a way of organizing two familiar activities into a more accountable interactional progression. The "accountable talk" framework provides a theoretical basis for why this sequence is logical: accountability to the learning community, to reasoning norms, and to new knowledge can reasonably be expected of students once they have had the opportunity to prepare their arguments (Hennessy et al., 2023). Furthermore, evidence from the fields of debate and oral argumentation is broadly consistent with the idea that a gradual rather than abrupt increase in interactional demands supports argument quality without overburdening students (Guo et al., 2023; González Palta et al., 2026), although neither study directly tested this staged TPS–Hot Seat sequence. The "accountability bridge" proposed here remains an interpretive concept drawn from this study's findings rather than a validated construct, and it should be examined in future designs that include a comparison group.

5.7 Limitations

Several limitations constrain the interpretation of these findings. First, the study involved a single class of 33 students without a comparison group; the results cannot establish comparative effectiveness or be statistically generalized to other schools. Second, the pre-intervention phase utilized Brown's five-indicator rubric (1–5 scale), whereas the action cycles employed a nine-indicator rubric (1–4 scale); consequently, the comparison between the pre-intervention phase and the cycles is contextual rather than a comparison of identical instruments. Third, Cycle II introduced multiple simultaneous changes—a narrower topic, altered timing, mandatory practice, language rules, targeted scaffolding, and audience incentives—making it impossible to isolate the independent contribution of each change. Fourth, although two raters assessed the work independently, only the moderation procedure and the reconciled final scores were recorded; the raw paired scores prior to reconciliation are unavailable, preventing a valid retrospective calculation of inter-rater reliability coefficients. Fifth, not all pair conversations were fully recorded, limiting detailed analysis of peer interactions. Sixth, there was no baseline measurement of the learning experience. Seventh, the qualitative analysis was interpretive and did not employ a formal thematic coding protocol involving multiple independent coders. Eighth, written consent documentation from students and parents/guardians—separate from general school permission—was not available. Ninth, potential researcher/teacher expectancy effects, the Hawthorne effect, and the impact of growing familiarity with the same assessment framework across cycles must be considered. Finally, some students still required direct scaffolding; therefore, the durability of their improvement following the withdrawal of such support remains unknown.

5.8 Directions for Future Research

Future research should examine the preparation-to-response sequence across multiple classes and schools and, where feasible, include comparison groups to strengthen causal inference. Studies should use the same assessment instrument at baseline and across cycles, retain paired raw scores from independent raters for reliability analysis, and record all Pair interactions to permit closer examination of peer talk. Further work could separate the contributions of TPS, Hot Seat, oral rehearsal, scaffolding, and audience accountability, and should examine whether improvements in public speaking performance are sustained after targeted teacher support is withdrawn. Because Diction and Grammar

remained unchanged in the present study, future interventions may also incorporate explicit and sustained language-form instruction alongside argumentation practice.

6. CONCLUSION

This classroom action research documented changes in the argumentative speaking performance and learning experiences of Grade XI students across two cycles of an integrated Think-Pair-Share and Hot Seat sequence. Mean speaking scores rose from 49.94 at the pre-action stage to 70.13 in Cycle I and 77.70 in Cycle II, with 30 out of 33 students (90.9%) reaching the threshold score of 75 by the end of Cycle II. Cross-cycle comparisons using the same instrument revealed the greatest improvements in Gestures and Expressions (+0.79) and Content Mastery (+0.58), while Diction and Grammar remained unchanged (2.91). Learning experience scores increased from 62.42 to 77.88 across the two cycles, with 29 out of 33 students (87.9%) meeting the threshold; the most significant gains occurred in Deep Learning Achievement, Self-Efficacy, and Active Interaction.

The primary implication is that the Hot Seat activity is most productive when students have already had the opportunity to construct and test their arguments rather than when spontaneity is demanded without preparation—a perspective that provides the Hot Seat with a clearer pedagogical rationale within argumentative speaking instruction. These findings indicate that the pedagogical value of this sequence lies in its progression from individual preparation to peer practice and then to responsive public argumentation, with the "Hot Seat" serving as a bridge of accountability between these stages. However, because this study employed a single-class CAR design lacking a control group, utilizing different baseline and cycle instruments, and introducing multiple changes simultaneously in Cycle II—the findings cannot be treated as controlled causal evidence that TPS and the Hot Seat independently produced the observed improvements. What the study offers is classroom-based evidence that a carefully designed preparation-to-response sequence can support the development of argumentative performance, while explicit and sustained linguistic practice remains necessary to improve diction and grammar.

DECLARATIONS

Funding

This research received no external funding.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgments

The authors gratefully acknowledge SMA Islam Pasuruan for providing the research setting and support, as well as the teachers and students who participated in this study.

7. REFERENCES

- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191–215. <https://doi.org/10.1037/0033-295X.84.2.191>
- Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Pearson Education.
- Brown, H. D. (2007). *Teaching by principles: An interactive approach to language pedagogy* (3rd ed.). Pearson Longman.
- Cooper, K. M., Schinske, J. N., & Tanner, K. D. (2021). Reconsidering the share of a Think–Pair–Share: Emerging limitations, alternatives, and opportunities for research. *CBE—Life Sciences Education*, 20(1), fe1. <https://doi.org/10.1187/cbe.20-08-0200>
- Dewey, J. (1997). *Experience and education*. Touchstone. (Original work published 1938)

- Fullan, M. (2013). *Stratosphere: Integrating technology, pedagogy, and change knowledge*. Pearson.
- Gillies, R. M. (2016). Cooperative learning: Review of research and practice. *Australian Journal of Teacher Education*, 41(3), 39–54. <https://doi.org/10.14221/ajte.2016v41n3.3>
- González Palta, I., Larrain, A., Contreras, R., & Guzmán, M. J. (2026). How do college students argue orally in classroom debates with different communicative goals? *Higher Education Quarterly*, 80(1), e70100. <https://doi.org/10.1111/hequ.70100>
- Guenther, A. R., & Abbott, C. M. (2024). Think-pair-share: Promoting equitable participation and in-depth discussion. *PRiMER*, 8, 7. <https://doi.org/10.22454/PRiMER.2024.444143>
- Guo, K., Zhong, Y., Li, D., & Chu, S. K. W. (2023). Effects of chatbot-assisted in-class debates on students' argumentation skills and task motivation. *Computers & Education*, 203, 104862. <https://doi.org/10.1016/j.compedu.2023.104862>
- Hall, J. K. (2022). L2 classroom interaction and its links to L2 learners' developing L2 linguistic repertoires: A research agenda. *Language Teaching*, 55(1), 100–115. <https://doi.org/10.1017/S0261444820000397>
- Hennessy, S., Calcagni, E., Leung, A., & Mercer, N. (2023). An analysis of the forms of teacher-student dialogue that are most productive for learning. *Language and Education*, 37(2), 186–211. <https://doi.org/10.1080/09500782.2021.2013252>
- Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, 70(2), 125–132. <https://doi.org/10.1111/j.1540-4781.1986.tb05256.x>
- Hu, L., & Chen, G. (2024). A systematic review and meta-analysis of productive peer talk moves. *Journal of Behavioral Education*, 33(4), 798–830. <https://doi.org/10.1007/s10864-023-09513-9>
- Kemmis, S., McTaggart, R., & Nixon, R. (2014). *The action research planner: Doing critical participatory action research*. Springer. <https://doi.org/10.1007/978-981-4560-67-2>
- Kolb, A. Y., & Kolb, D. A. (2017). Experiential learning theory as a guide for experiential educators in higher education. *Experiential Learning & Teaching in Higher Education*, 1(1), 7–44. <https://doi.org/10.46787/elthe.v1i1.3362>
- Nadolski, R. J., Hummel, H. G. K., Rusman, E., & Ackermans, K. (2021). Rubric formats for the formative assessment of oral presentation skills acquisition in secondary education. *Educational Technology Research and Development*, 69(5), 2663–2682. <https://doi.org/10.1007/s11423-021-10030-7>
- Padmawati, K. D., Arini, N. W., & Yudiana, K. (2019). Analisis keterampilan berbicara siswa kelas V pada mata pelajaran Bahasa Indonesia. *Journal for Lesson and Learning Studies*, 2(2), 190–200. <https://doi.org/10.23887/jlls.v2i2.18626>
- Raba, A. A. A. (2017). The influence of Think-Pair-Share (TPS) on improving students' oral communication skills in EFL classrooms. *Creative Education*, 8, 12–23. <https://doi.org/10.4236/ce.2017.81002>
- Sybing, R. (2021). Examining dialogic opportunities in teacher-student interaction: An ethnographic observation of the language classroom. *Learning, Culture and Social Interaction*, 28, 100492. <https://doi.org/10.1016/j.lcsi.2021.100492>
- Toulmin, S. E. (2003). *The uses of argument* (Updated ed.). Cambridge University Press.
- Van der Veen, C., Michaels, S., Dobber, M., van Kruistum, C., & van Oers, B. (2021). Design, implementation, and evaluation of dialogic classroom talk in early childhood education. *Learning, Culture and Social Interaction*, 29, 100515. <https://doi.org/10.1016/j.lcsi.2021.100515>
- Van der Wilt, F., Bouwer, R., & van der Veen, C. (2022). Dialogic classroom talk in early childhood education: The effect on language skills and social competence. *Learning and Instruction*, 77, 101522. <https://doi.org/10.1016/j.learninstruc.2021.101522>
- Veenstra, R. J., & Weaver, C. (2022). Examining a continuum of FL speaking anxiety over time in an EFL classroom in Japan. *System*, 110, 102889. <https://doi.org/10.1016/j.system.2022.102889>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Yu, Q. (2024). Foreign language anxiety research in System between 2004 and 2023: Looking back and looking forward. *Frontiers in Psychology*, 15, 1373290. <https://doi.org/10.3389/fpsyg.2024.1373290>
- Zarefsky, D. (2009). *Argumentation: The study of effective reasoning* (3rd ed.). The Teaching Company.

- Zhang, X. (2019). Foreign language anxiety and foreign language performance: A meta-analysis. *The Modern Language Journal*, 103(4), 763–781. <https://doi.org/10.1111/modl.12590>
- Zhang, X., & Ardasheva, Y. (2019). Sources of college EFL learners' self-efficacy in the English public speaking domain. *English for Specific Purposes*, 53, 47–59. <https://doi.org/10.1016/j.esp.2018.09.004>
- Zhang, X., Ardasheva, Y., & Austin, B. W. (2020). Self-efficacy and English public speaking performance: A mixed method approach. *English for Specific Purposes*, 59, 1–16. <https://doi.org/10.1016/j.esp.2020.02.001>