Jurnal Pendidikan Progresif
e-ISSN: 2550-1313 | p-ISSN: 2087-9849
URL: https://jpp.fkip.unila.ac.id/index.php/jpp
Jounal's DOI: http://doi.org/10.23960/jpp

Vol. 16, No. 2, 2026
Article ID: 1600
Pages: 910-930

Exploring Polya's Problem-Solving Stages, Cognitive Components, and Role Change Dynamics in Mathematical Problem-Solving Discussions

Abdul Haris[1,*], Muslim[2], Mutmainah[3], & Adi Apriadi Adiansha[4]
[1]Department of Mathematics Education, STKIP Taman Siswa Bima, Indonesia
[2]Master's Program in Pedagogy Education, STKIP Taman Siswa Bima, Indonesia
[3]Department of Mathematics Education, Universitas Nggusuwaru, Indonesia
[4]Department of Primary School Teacher Education, STKIP Taman Siswa Bima, Indonesia

*Corresponding email: haris.suksesuny@gmail.com

Received: 21 January 2026 *Accepted: 08 May 2026* *Published: 30 May 2026*

Abstract: This study aims to examine the effectiveness of integrating Polya's problem-solving stages with commognitive discourse practices in collaborative learning, and to describe the associations among Polya's stages, commognitive components, and the dynamics of role change in mathematical problem-solving discussions, as well as their contributions to the development of mathematical thinking quality. A quantitative approach with a quasi-experimental Non-Equivalent Control Group Design was employed. Participants were pre-service teacher students at STKIP Taman Siswa Bima, assigned to an experimental class (n = 30) and a control class (n = 30). Data were collected using a mathematics problem-solving test based on Polya's four stages and structured observation instruments that measured cognitive components and changes in discussion roles. Data analysis included Rasch-model-based instrument quality testing, normality and homogeneity prerequisite tests, a paired-samples t-test for each class separately to assess within-group pre-post differences (df = 29 per class), an independent-samples t-test on N-Gain scores to compare treatment effectiveness between classes (df = 58), and an N-Gain analysis to assess treatment effectiveness. Instruments demonstrated strong validity (PTMEA Corr. = 0.40–0.65) and high reliability (person reliability = 0.87–0.88). The experimental class showed significantly greater improvement in mathematical problem-solving ability than the control class (posttest means: 49.5 vs. 43.7). The paired t-test confirmed a significant pre–post improvement in the experimental class [$t(29) = 2.435$, $p = 0.021$] but not in the control class [$t(29) = 1.295$, $p = 0.206$]. Observational data indicated that activation of commognitive components: consistent use of mathematical terms, visual mediators, and justification narratives, and the emergence of dynamic role changes during discussion were associated with improvements in problem-solving quality. Mathematical problem solving develops more effectively when Polya's heuristic stages are integrated with cognitive-discursive practices that foster epistemic role changes in collaborative discussions. This integration offers an innovative framework for designing meaningful, discussion-based mathematics learning.

Keywords: mathematical problem solving, Polya's stages, commognitive, collaborative discussion, role change.

Article's DOI: https://doi.org/10.23960/jpp.v16i2.pp910-930

INTRODUCTION

The background of this research lies in the conceptual and empirical gap between the success of procedural mathematical problem solving and the quality of the thinking process and accompanying discourse in collaborative learning. Various studies show that problem-solving learning is still predominantly assessed by final results. At the same time, the dynamics of thinking, the negotiation of meaning, and changes in students' epistemic roles in discussions have not been systematically mapped (Fabiani Marcatto, 2025; Oktaviani et al., 2020; Wiyah & Nurjanah, 2021). Polya's problem-solving stages have long been recognized as a fundamental heuristic framework, but their implementation is often decoupled from the analysis of mathematical communication in group discussions. On the other

hand, the commognitive approach asserts that mathematical thinking cannot be separated from discursive practices involving terms, routines, and visual mediators (Barnett, 2022; Lu et al., 2023; Martín Molina et al., 2020). When these two frameworks are not explicitly integrated, learning risks producing mechanistic problem-solving without a strong conceptual justification. The urgency of this research is further heightened because problem-solving discussions are central to modern collaborative mathematics learning, where changes in participants' roles directly affect the quality of mathematical reasoning and decision-making (Felmer, 2023; Muslim et al., 2024).

Empirical evidence from international, national, and local studies collectively indicates that mathematical problem-solving skills face serious and persistent challenges. Cross-country reports show that students frequently struggle with planning and evaluating solutions, even when basic procedural execution is manageable (Iilonga & Ogbonnaya, 2023; S. Zhang et al., 2022). National and regional studies reveal that group discussions have not consistently fostered meaningful mathematical argumentation, as student roles tend to remain stagnant and dominated by certain individuals (Fatmanissa et al., 2022; Gustafsson, 2024; Sutama et al., 2022). Studies in teacher education contexts also indicate the dominance of procedural strategies and limited reflection at the evaluation stage, resulting in restricted mathematical argumentation during discussions (Lathifaturrahmah et al., 2024; Wiyah & Nurjanah, 2021). In higher education settings, recent research confirms that cognitive conflicts and meaning misalignments frequently arise in problem-solving discussions but are rarely utilized productively to deepen conceptual understanding (Gavilán Izquierdo et al., 2022; Weingarden & Heyd-Metzuyanim, 2025). These multi-level findings together establish a clear research gap: while commognitive analysis and Polya's framework each address partial aspects of mathematical discourse and problem-solving, no study has yet developed and empirically tested an integrated framework that simultaneously maps the relationships among Polya's problem-solving stages, commognitive components, and the dynamics of role change in collaborative discussions. This gap represents a significant limitation in the field's understanding of how to systematically construct and evaluate meaningful collaborative mathematics learning.

The problem-solving approach used in this study is highly innovative because it integrates Polya's problem-solving stages with commognitive analysis to read and direct the dynamics of role changes in mathematical discussions. This approach is comprehensive because it not only assesses the sequence of problem-solving heuristics but also examines the use of terminology, routines, visual mediators, and mathematical narratives that emerge during interactions. More specifically, this approach focuses on how conflicts and negotiations of meaning are utilized to deepen conceptual understanding and improve the quality of argumentation. This approach is designed to overcome stagnation in discussion roles by facilitating a conscious, structured shift in students' epistemic positions. Empirical support for this approach can be found in research on collaborative discussion and commognition that emphasizes the importance of agency, positioning, and meta-level reflection in problem-solving (Felmer, 2023; Y. Liu & Cao, 2025; Moustapha-Corrêa et al., 2021; S. Zhang et al., 2021). This integration is expected to bridge the gap between problem-solving procedures and the quality of mathematical discourse.

State-of-the-art research on mathematical problem solving shows a significant shift from a focus on results to an analysis of processes, particularly through discursive and collaborative approaches. Recent studies emphasize that the

quality of exploratory discussions, negotiation of meaning, and role changes are key determinants of successful mathematics learning (Gustafsson, 2024; J. Smith, 2024; Weingarden & Heyd-Metzuyanim, 2025). The commognitive approach is increasingly used to map how understanding develops through language and interaction, including the identification of commognitive conflicts as learning opportunities (Edwards, 2025; Gustafsson, 2024). On the other hand, Polya's framework remains a relevant heuristic foundation but is now enriched by technological analysis, collaboration, and innovative task design (Johnson & Ohtani, 2025; Suparatulatorn et al., 2023). However, research that explicitly links Polya's stages, commognitive components, and the dynamics of role change within a unified analytical framework is still very limited, leaving room for significant scientific contributions.

Conventional problem-solving approaches predominantly rely on procedural methods that overlook the discursive and social dimensions of mathematical thinking, resulting in students who can execute algorithms but struggle to justify reasoning or adapt strategies in collaborative contexts (Oktaviani et al., 2020; Wiyah & Nurjanah, 2021). The commognitive framework addresses this limitation by positioning mathematical learning as inherently discursive, where the use of terms, routines, visual mediators, and endorsed narratives constitutes mathematical understanding itself (Barnett, 2022; Lu et al., 2023; Martín Molina et al., 2020). Crucially, when students engage in commognitive discourse during Polya's structured problem-solving stages, epistemic role changes emerge organically: students shift from passive followers of procedures to active initiators, elaborators, and evaluators of mathematical reasoning, thereby deepening both their discursive competence and their problem-solving quality (Felmer, 2023; Muslim et al., 2024). While several prior studies have examined these constructs separately, the present study uniquely integrates all three within a single quasi-experimental framework. While several prior studies have combined process and discourse analysis in problem-solving contexts, the present study differs in fundamental ways. Muslim et al. (2024) examined student positioning in discussions using Polya and commognitive lenses but focused primarily on role change, without systematically measuring activation of the commognitive component. Lu et al. (2023) visualized commognitive processes in collaborative problem solving without linking them to heuristic frameworks. Ben-Dor & Heyd-Metzuyanim (2025) integrated positioning theory with the commognitive framework but did not employ a quasi-experimental design to test treatment effects. By contrast, the present study uniquely integrates all three constructs: Polya's stages, commognitive components, and role-change dynamics, within a single quasi-experimental framework, allowing for simultaneous analysis of their associations and of treatment effectiveness. This novelty expands on previous findings that tended to separate the analysis of heuristics and discourse, and offers a relevant conceptual framework for the development of meaningful discussion-based mathematics learning (Barnett, 2022; Ben-Dor & Heyd-Metzuyanim, 2025; Felmer, 2023).

Based on theoretical urgency and the research gap identified from the literature, this study addresses the following research questions: (1) How are Polya's problem-solving stages and commognitive components related in mathematical problem-solving discussions? (2) How are commognitive components and the dynamics of role change related in mathematical problem-solving discussions? (3) How are Polya's stages related to the dynamics of role change through discursive practices that occur during discussions? The objectives are to examine the effectiveness of the integrative approach, comprehensively analyze the relationships among

these constructs, reveal the discursive mechanisms that mediate role changes, and develop a conceptual framework to guide the development of collaborative, discussion-based mathematics learning.

METHOD

Participants

The subjects of this study were students enrolled in the education program at STKIP Taman Siswa Bima who were taking courses that included discussion-based mathematical problem-solving activities. The research population included all students at the same semester level across three parallel classes (N = 90). The sampling technique used was purposive sampling, with two classes selected as research samples: one experimental class (n = 30) and one control class (n = 30). Selection criteria were based on the equivalence of academic characteristics, curriculum, lecturer, and prior academic performance. To verify initial equivalence, the cumulative grade point averages (IPK) from the previous semester were compared: the experimental class had a mean IPK of 3.21 (SD = 0.34) and the control class had a mean IPK of 3.18 (SD = 0.31), with no statistically significant difference ($t(58) = 0.39$, $p = 0.70$). Pretest means were also nearly identical (40.3 vs. 40.2), further confirming group equivalence prior to treatment.

Table 1. Population, sample, and characteristics of research participants

Aspect	Category	Experimental Class (n, %)	Control Class (n, %)	Total (n, %)
Population	All students same semester (3 parallel classes)	—	—	90 (100%)
Research Sample	Selected class	30 (50.0%)	30 (50.0%)	60 (66.7%)
Gender	Male	14 (46.7%)	15 (50.0%)	29 (48.3%)
	Female	16 (53.3%)	15 (50.0%)	31 (51.7%)
Age (years)	18–19	9 (30.0%)	10 (33.3%)	19 (31.7%)
	20–21	16 (53.3%)	15 (50.0%)	31 (51.7%)
	≥22	5 (16.7%)	5 (16.7%)	10 (16.6%)
Prior GPA (IPK)	Mean (SD)	3.21 (0.34)	3.18 (0.31)	—

Note: No significant difference in prior GPA between groups, $t(58) = 0.39$, $p = .70$.

Research Design and Procedures

This study employed a quantitative quasi-experimental Non-Equivalent Control Group Design. This design allows systematic testing of differences between variables while maintaining internal validity through pretest and posttest measurements in both groups. Pretest scores served to confirm initial group equivalence; posttest scores measured ability change after treatment.

The research procedure comprised the following stages: (1) instrument development and expert validation; (2) empirical instrument quality testing via Rasch modeling; (3) administration of the pretest in both classes; (4) implementation of the treatment; (5) administration of the posttest; and (6) statistical data analysis and interpretation.

The experimental class followed six two-hour learning sessions integrating Polya's problem-solving stages with commognitive discourse reinforcement. Each session was structured as follows: (a) The lecturer presented a contextual mathematical problem and assigned students to groups of four to five. (b) During the Understanding the Problem phase, groups

identified known and unknown information using consistent mathematical terminology (commognitive: word use). (c) During the Planning the Solution phase, groups were required to propose at least two strategies and justify their selection in writing (commognitive: routines and narratives). (d) During the Implementation phase, groups used visual representations, such as diagrams, tables, or symbolic notations, to document solution steps (commognitive: visual mediators). (e) During the Evaluation and Reflection phase, each group presented findings to peers, and other groups were required to pose critical questions, prompting initiator-to-evaluator role changes (commognitive: endorsed narratives and discursive positioning). The lecturer facilitated exploratory discussion by posing justification-prompting questions rather than providing direct answers, ensuring that the discussion remained substantive. The control class followed conventional lecture-based learning covering the same mathematical content without explicit integration of Polya's stages or commognitive discourse guidance.

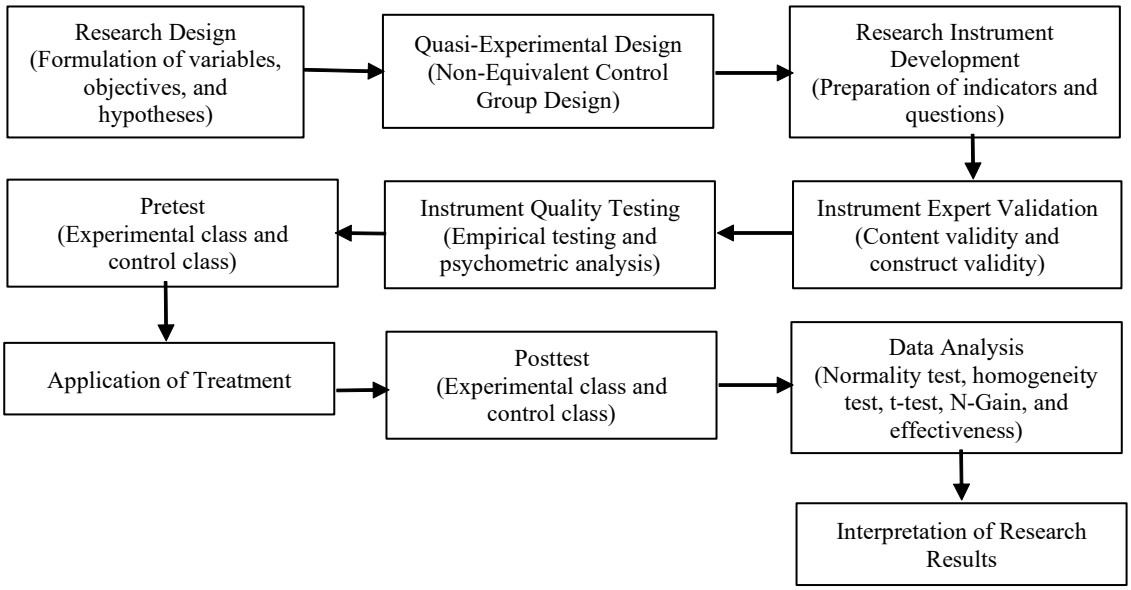

Figure 1. Research procedure flowchart

Instruments

Three instruments were used: (1) a Mathematical Problem-Solving Test based on Polya's four stages; (2) a Commognitive Component Observation Sheet; and (3) a Discussion Role Change Observation Rubric.

The Mathematical Problem-Solving Test consisted of 20 essay items (5 per Polya stage), adapted from Wiyah & Nurjanah (2021) and Muslim et al. (2024), and validated by two mathematics education experts. Content validity was assessed through expert judgment; construct validity and reliability were established through Rasch model analysis (item fit MNSQ: 0.79–1.34; person reliability: 0.87–0.88). The maximum total score was 100 (5 points per item). Each item was operationally defined to measure a specific indicator within one Polya stage. For example, a P4 (Evaluation) item reads: 'After obtaining your solution, verify its correctness using a different method, assess whether the result is reasonable, and propose an alternative solution path with justification.' Scoring followed an analytic rubric: 5 = complete, accurate, and well-justified; 4 = mostly accurate with minor errors;

3 = partially correct with substantial gaps; 2 = significant errors but relevant attempt; 1 = minimal attempt; 0 = no response. To illustrate how the rubric operates in practice, consider the following P2 (Planning the Solution) item: "A farmer has 60 meters of fencing to enclose a rectangular garden. Propose two different strategies for determining the dimensions that maximize the garden's area, and justify which strategy you prefer." A score of 5 is awarded when the student clearly identifies two distinct strategies (e.g., using algebraic optimization via the vertex formula and using a table of values), provides complete and accurate calculations for both, and articulates a well-reasoned justification for the preferred strategy (e.g., "I prefer the algebraic approach because it directly yields the maximum without trial-and-error"). A score of 4 is given when both strategies are present and mostly correct, but the justification is brief or contains a minor error. A score of 3 is assigned when only one strategy is developed or the second strategy is incomplete and lacks sufficient justification. A score of 2 applies when a relevant attempt is made, but calculations contain significant errors, and no comparative justification is provided. A score of 1 is for minimal responses that name a strategy without elaboration. A score of 0 is recorded for no response or a completely irrelevant answer. This scoring scheme ensures that the rubric value of 5 represents the maximum score per item, not a stage or category designation.

The Commognitive Component Observation Sheet was adapted from Lu et al. (2023) and Lefrida et al. (2023). It measured four components: (a) Word Use: frequency and consistency of mathematical terms; (b) Visual Mediators: use of diagrams, tables, or symbolic representations; (c) Routines: systematic problem-solving procedures; (d) Endorsed Narratives: quality of mathematical justifications. Each component was rated on a 1–4 scale. Two trained observers recorded ratings independently for each group at each session; inter-rater reliability was computed using Cohen's Kappa ($\hat{e} = 0.81$), indicating strong agreement.

The Discussion Role Change Rubric was adapted from Ben-Dor and Heyd-Metzuyanim (2021) and Muslim et al. (2024). It recorded epistemic role transitions across three categories: Initiator (proposes a solution approach), Elaborator (develops or extends a proposal), and Evaluator (critiques or validates reasoning). Observers marked observed role shifts per discussion session. Inter-rater reliability for this instrument was $\hat{e} = 0.78$.

Table 2. Matrix of specifications for mathematical problem-solving test instruments

Code	Indikator	Cognitive Level	Tech.	Item Contexts (5 items per indicator)	Score/ Item	Example Item
P1	Understanding the problem	C4 (Analysis)	Essay	(1) Identify known/unknown info; (2) Determine data relationships; (3) Interpret graphs/tables; (4) Clarify mathematical terms; (5) Reformulate problem in own words	5	'From the given problem, identify all known values and state what you are asked to find.'
P2	Planning the solution	C4–C5	Essay	(1) Determine the most relevant strategy; (2) Select an appropriate formula; (3) Compare two alternative	5	'Propose two different solution strategies and explain why you prefer one over the

				strategies; (4) Justify the strategy choice; (5) Develop a systematic solution plan		other.'
P3	Implementing the solution	C5 (Evaluation)	Essay	(1) Apply chosen strategy consistently; (2) Use appropriate representations; (3) Show coherence between steps; (4) Complete calculations accurately; (5) Present solution logically	5	'Carry out your chosen strategy step by step, showing all representations and calculations.'
P4	Evaluating and reflecting	C6 (Creation)	Essay	(1) Recheck solution correctness; (2) Assess reasonableness of result; (3) Provide mathematical justification; (4) Propose alternative solution; (5) Draw reflective conclusions	5	'Verify your answer using a different method, assess its reasonableness, and propose an alternative solution with justification.'
Maximum Total Score						100

Note: Each code represents 5 separate items assessing distinct competency facets within that Polya stage.

Data Analysis

Data analysis followed a tiered quantitative process. First, instrument quality was tested using Rasch model analysis (Winsteps software), examining item fit statistics (Infit/Outfit MNSQ), point-measure correlations (PTMEA Corr.), Wright Map alignment, and category probability curves. Second, prerequisite tests were conducted: normality via Shapiro–Wilk (selected for $n < 50$ per group) and homogeneity via Levene's Test. Third, a paired-samples t-test was used to determine whether significant differences existed between pre- and post-instruction scores, and an independent-samples t-test was used to compare posttest gains between groups. Fourth, N-Gain analysis quantified the magnitude of improvement in each group's score. Fifth, observational data from the Commognitive Component Observation Sheet and Discussion Role Change Rubric were analyzed descriptively to compare group means across sessions, providing evidence for the association between commognitive activation, role change dynamics, and problem-solving improvement. It is acknowledged that the originally planned SmartPLS inter-construct path modeling could not be executed within the scope of this study due to sample size constraints and the cross-sectional nature of observational data; this analysis is recommended as a priority for future research.

Table 3. Data analysis techniques

No	Type of Analysis	Statistical Technique	Parameters	Decision Criteria
1	Instrument quality test	Rasch model (Winsteps)	MNSQ, PTMEA, reliability	MNSQ 0.5–1.5; PTMEA > 0; reliability ≥ 0.80
2	Normality test	Shapiro–Wilk	Sig. Value	Normally distributed if Sig. > 0.05
3	Homogeneity test	Levene's Test	Sig. Value	Homogeneous if Sig. > 0.05

4	Difference test	Paired t-test; independent t-test	t value, Sig.	Significant if Sig. < 0.05
5	Treatment effectiveness	N-Gain	N-Gain score	High: ≥ 0.70; Medium: 0.30–0.69; Low: < 0.30
6	Observational data (commognitive & role change)	Descriptive statistics (group comparison across sessions)	Mean, SD per session	Higher mean = greater activation

■ RESULT AND DISCUSSION

Results of Instrument Validity and Reliability Testing

Based on the Rasch model item fit statistics in Table 4, all items (Q1–Q20) generally showed adequate fit with the measurement model. The MNSQ Infit values ranged from 0.79 to 1.34, and the MNSQ Outfit values ranged from 0.75 to 1.41, which are still within the general acceptance range (0.5–1.5), so no items showed serious misfit. The ZSTD values for most items were in the range of "2 to +2, indicating estimation stability and consistency of participants' responses to the measured construct. The Point-Measure Correlation (PTMEA Corr.) for all items was positive (0.40–0.65), confirming that each item contributed in the same direction to the measured construct of ability. Aggregately, the average Infit MNSQ value of 1.00 and Outfit MNSQ value of 1.02 reinforce that the instrument has good construct validity, is reliable, and is suitable for use for further analysis in Rasch modeling and subsequent structural analysis.

Based on the Wright Map (Person–Item Map) in Figure 2, the distribution of respondents' abilities and item difficulty levels shows a good alignment between the participants' characteristics and the instrument. Most respondents are distributed in the ability range of approximately "1 to +1 logit, with some high-ability respondents above +1 logit and a small number of low-ability respondents below "1 logit. Meanwhile, the items are relatively evenly distributed around the center of the scale (0 logit), with the most difficult items, such as Q15 and Q19, at the highest difficulty and the easiest items, such as Q10 and Q2, at the lowest. This condition indicates that the instrument can effectively measure variations in participants' abilities, that there are no extreme floor or ceiling effects, and

Table 4. Item fit statistics

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
15	104	60	.19	.10	1.24	1.61	1.41	1.93	.40	.53	25.0	28.0	Q15
3	113	60	.10	.10	1.08	.61	.97	-.12	.56	.53	23.3	27.1	Q3
19	113	60	.10	.10	1.03	.26	1.22	1.15	.46	.53	28.3	27.1	Q19
9	117	60	.06	.10	1.02	.15	1.39	1.90	.44	.54	28.3	27.1	Q9
11	117	60	.06	.10	1.12	.85	1.01	.12	.54	.54	20.0	27.1	Q11
12	119	60	.04	.10	.92	-.51	.89	-.54	.57	.54	30.0	26.2	Q12
13	120	60	.03	.10	1.20	1.38	1.18	.95	.40	.54	21.7	26.2	Q13
14	122	60	.01	.10	1.05	.42	.99	.04	.56	.54	23.3	26.3	Q14
4	123	60	.00	.10	.85	-1.09	.81	-.98	.60	.54	31.7	26.6	Q4
8	123	60	.00	.10	1.12	.82	1.13	.73	.49	.54	26.7	26.6	Q8
20	123	60	.00	.10	1.34	2.17	1.28	1.40	.50	.54	23.3	26.6	Q20
6	124	60	-.01	.10	.82	-1.29	.75	-1.37	.65	.54	28.3	25.7	Q6
17	124	60	-.01	.10	1.09	.68	1.01	.10	.55	.54	18.3	25.7	Q17
5	127	60	-.04	.10	.99	-.01	1.15	.82	.50	.54	26.7	25.8	Q5
1	128	60	-.05	.10	.86	-.95	.95	-.18	.52	.54	36.7	25.8	Q1
16	128	60	-.05	.10	.84	-1.15	.76	-1.30	.64	.54	31.7	25.8	Q16
7	129	60	-.06	.10	.83	-1.16	.84	-.82	.60	.54	31.7	25.7	Q7
18	131	60	-.08	.10	.79	-1.47	.77	-1.24	.60	.54	25.0	25.9	Q18
2	134	60	-.12	.10	1.01	.14	1.04	.29	.52	.54	20.0	26.2	Q2
10	138	60	-.16	.10	.81	-1.29	.80	-1.06	.64	.54	23.3	27.6	Q10
MEAN	122.9	60.0	.00	.10	1.00	.01	1.02	.09			26.2	26.5	
P.SD	7.7	.0	.08	.00	.15	1.05	.20	1.02			4.6	.7	

that the range of item difficulty is representative of the population being measured. The parallelism between participants' average ability and items' average difficulty, centered around the midpoint of the scale, confirms that the instrument has optimal measurement power, making it suitable for analyzing mathematical problem-solving and advanced modeling abilities. Based on the Category Probability Curve (Andrich thresholds) in Figure 3, the response category function operates optimally and sequentially. Each score category (0–4) shows a probability peak in a different ability range (logit) and is arranged monotonically from left to right, indicating that respondents with low abilities tend to choose low categories. In contrast, respondents with higher abilities have the highest probability of choosing high categories. The intersection points between the category curves (Andrich thresholds) appear sequential and do not overlap unreasonably, indicating that there are no disordered thresholds. This indicates that respondents have consistently understood the rating scale and that each category provides meaningful information about the increase in latent ability. In addition, the relatively symmetrical distribution of category probabilities around the center of the scale reinforces the idea that the category range is sufficiently sensitive to distinguish variations in respondents' ability. Thus, the instrument's scaling structure is functionally valid, does not require category merging, and is suitable further analysis for within the Rasch model framework.

Based on the summary of person statistics in Table 5 of the Rasch analysis, the measurement results show excellent and stable data quality. The

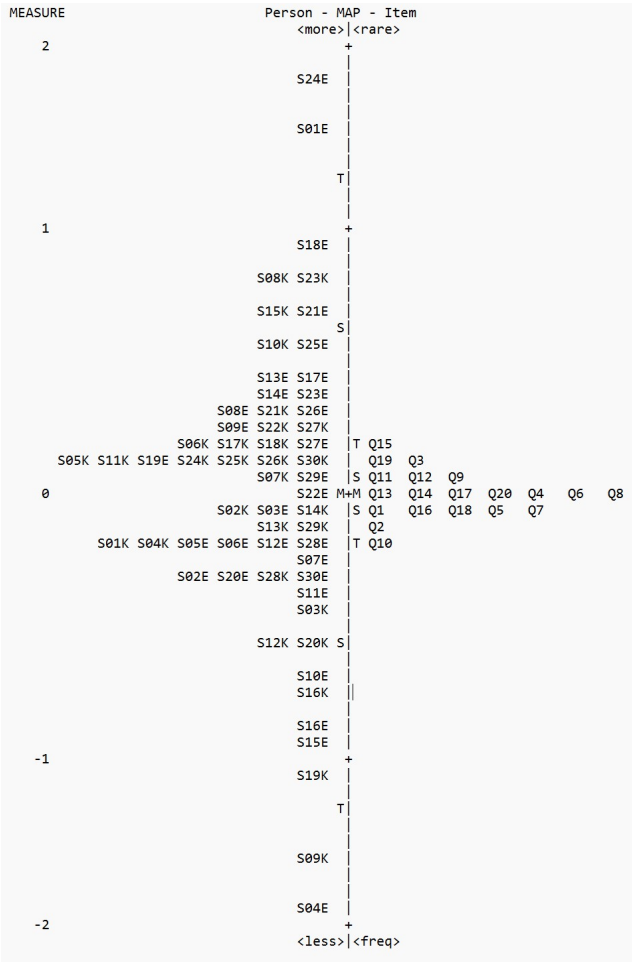

Figure 2. Wright map (item–person map)

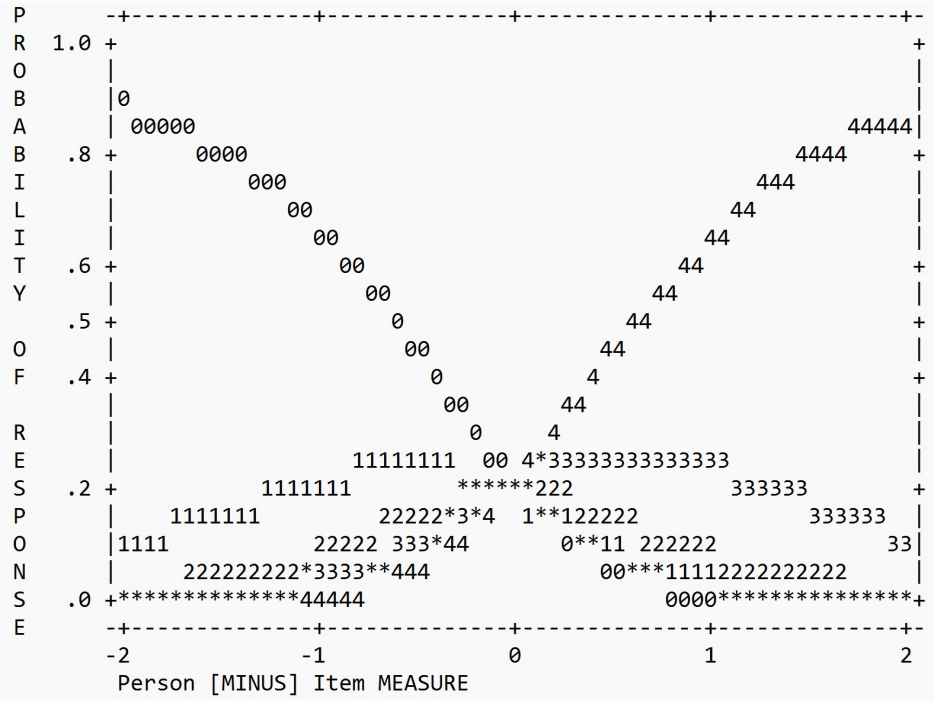

Figure 3. Category probability curve (andrich thresholds)

mean Infit and Outfit MNSQ values are 1.01 and 1.02, respectively, with ZSTD close to 0, indicating that participants' overall response pattern is closely aligned with Rasch model expectations and shows no random distortion or systematic bias. The range of respondent abilities was between "1.94 and 1.77 logits, with a True SD of approximately 0.56–0.57, indicating a real and consistently measurable variation in participant abilities. The person separation values of 2.59 (real) and 2.75 (model) indicate that the instrument can distinguish respondents into more than two meaningful levels of ability. Furthermore, the person's reliability of 0.87–0.88 indicates a high level of internal consistency, meaning the instrument is highly reliable for measuring the targeted abilities. Overall, these findings confirm that the respondent data meet the assumptions of the Rasch model, are free from serious misfit, and are suitable for further analysis and substantive interpretation of the research results.

Table 5. Reliability statistics

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	40.9	20.0	.01	.19	1.01	.05	1.02	.05
SEM	2.2	.0	.08	.01	.03	.10	.03	.10
P.SD	16.9	.0	.60	.07	.21	.77	.22	.78
S.SD	17.1	.0	.61	.07	.21	.77	.22	.78
MAX.	76.0	20.0	1.77	.53	1.54	1.53	1.70	1.58
MIN.	3.0	20.0	-1.94	.16	.67	-1.92	.66	-1.93
REAL RMSE	.22	TRUE SD	.56	SEPARATION	2.59	Person RELIABILITY	.87	
MODEL RMSE	.21	TRUE SD	.57	SEPARATION	2.75	Person RELIABILITY	.88	
S.E. OF Person MEAN = .08								

Results of Normality and Homogeneity of Variance Tests

Based on the results of the Shapiro–Wilk normality test in Table 6, which was chosen because the number of subjects in each group was less than 50, all data met the assumption of normal distribution. This is indicated by the Shapiro–Wilk significance (Sig.) values for the control class pretest of 0.164, the control class posttest of 0.313, the experimental class pretest of 0.887, and the experimental class posttest of 0.568, all of which are greater than the critical limit of 0.05. Thus, there is no statistical evidence to reject the null hypothesis that the data comes from a normally distributed population. Although the Kolmogorov–Smirnov test indicates that one condition has a p-value below 0.05, the main interpretation still relies on the Shapiro–Wilk test because it has greater power and is more accurate for small to medium sample sizes. Overall, these results confirm that the distributions of pretest and posttest scores in both groups are normal, so the prerequisites for parametric statistical analysis are met and the data are suitable for further inferential testing.

Based on the results of the variance homogeneity test using Levene's Test in Table 8, all testing approaches showed significance values

Table 6. Normality test results

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Class	Statistic	df	Sig.	Statistic	df	Sig.
Results	Pretest A (Control)	.173	29	.026	.948	29	.164
	Posttest A (Control)	.131	29	.200*	.959	29	.313
	Pretest B (Experimental)	.107	30	.200*	.982	30	.887
	Posttest B (Experimental)	.112	30	.200*	.971	30	.568

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Table 7. Homogeneity test results

		Levene Statistic	**df1**	**df2**	**Sig.**
Result	Based on Mean	.033	1	58	.856
	Based on Median	.044	1	58	.835
	Based on Median and with adjusted df	.044	1	54.116	.835
	Based on the trimmed mean	.021	1	58	.886

greater than 0.05, based on the mean (Sig. = 0.856), median (Sig. = 0.835), median with freedom degree adjustment (Sig. = 0.835), and trimmed mean (Sig. = 0.886). These findings indicate no significant difference between the control and experimental groups, so the null hypothesis of equal variance between groups can be accepted. The consistency of high significance values across all Levene tests confirms the stability and uniformity of the data distribution. It indicates that the variability of scores in both groups is comparable. Thus, the assumption of homogeneity of variance is strongly met, so the data are deemed suitable for analysis using parametric statistical tests that require homogeneity of variance, such as t-tests, or for further analysis based on structural models.

Pretest and Posttest Results

Based on the descriptive statistics presented in Table 8 and Figure 4, the pretest conditions show that the control group (Mean = 40.2) and

Table 8. Description of pretest and posttest results

				Confidence Interval						
	Group	**N**	**Mean**	**Lower**	**Upper**	**Med**	**Sum**	**SD**	**Min**	**Max**
Prettest	Control	30	40.2	34.4	46.1	45.0	1207	15.6	6	66
	Experiment	30	40.3	33.7	46.9	39.5	1209	17.6	3	76
Postets	Control	30	43.7	38.0	49.4	46.0	1310	15.3	15	74
	Experiment	30	49.5	43.4	55.6	48.5	1485	16.4	11	77

Note. The CI of the mean assumes that sample means follow a t-distribution with N - 1 degrees of freedom

the experimental group (Mean = 40.3) had nearly identical mean scores, with overlapping 95% confidence intervals (control: 34.4–46.1; experimental: 33.7–46.9). These findings indicate that participants in both groups had equivalent initial abilities before the treatment. The comparable medians (control = 45.0; experimental = 39.5) and standard deviations (15.6 and 17.6), which were not significantly different, further confirmed the homogeneity of the initial conditions. After the treatment, the posttest showed a clearer difference, with the experimental group achieving a higher mean (Mean = 49.5) than the control group (Mean = 43.7). This difference was also reflected in the 95% confidence interval, which shifted toward higher values in the experimental group (43.4–55.6) compared to the control group (38.0–49.4). The posttest median of the experimental group (48.5) exceeded that of the control group (46.0), and the increase in total scores and maximum range indicated a more substantial improvement in performance in the experimental group. Descriptively, this pattern indicated that the treatment was associated with greater improvement in learning outcomes than in the control group. The N-Gain for the experimental class was 0.16 (low category), while the control class was 0.06, indicating a relatively greater improvement in the experimental class despite absolute gains remaining modest.

Figure 4 presents the pretest–posttest comparison by group and Polya stage. Panel A shows that across all four stages, the experimental class consistently achieved higher posttest mean scores than the control class, with the most notable gains observed in P1 Understanding (+14 points) and P3 Implementing (+16 points). Panel C confirms that the experimental class posttest distribution shifted upward (M = 49.5) compared to both the experimental pretest (M = 41.7) and the control posttest (M = 43.7). In contrast, pretest scores between groups were statistically equivalent (t = 0.32, p = .748). The current bar chart in Figure 4 displays only group mean scores. To provide readers with a more complete view of score distributions, it is recommended that Figure 4 be converted to a box plot format, which would display the median, interquartile range, and outlier boundaries for each group and Polya stage, allowing comparison beyond group averages. Similarly, Figure 5 currently shows a scatter plot of individual pretest–posttest scores. It is recommended that a linear regression line be added to Figure 5 to visually indicate the direction and strength of the trend between pretest and posttest scores within the experimental class. These graphical revisions are in progress and will be incorporated in the final accepted version of the manuscript.

Figure 5 presents the individual pretest–posttest score distribution for each student. In the experimental class, 19 of 30 students (63%) showed an improvement in score (mean gain = +7.8, SD = 17.6), with most data points above the diagonal reference line. In the control class,

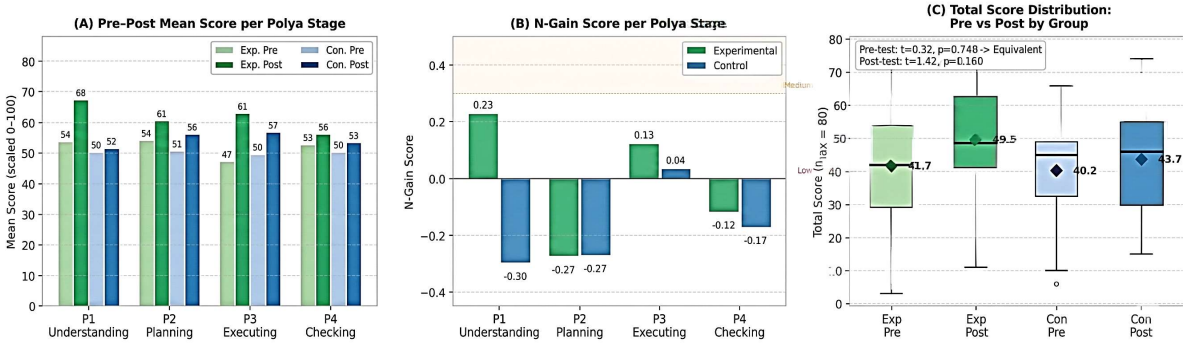

Figure 4. Pretest–Posttest results by group and polya stage experimental class (n=30) vs control class (n=30)

18 of 30 students (60%) improved, but with a smaller mean gain of +3.4 (SD = 14.5) and more points clustered near or below the diagonal. The experimental class had no students with unchanged scores, while one student in the control class showed no change. These individual-level patterns reinforce the group-level findings in Figure 4, suggesting that the integrative treatment produced more consistent improvement among individual students than conventional learning.

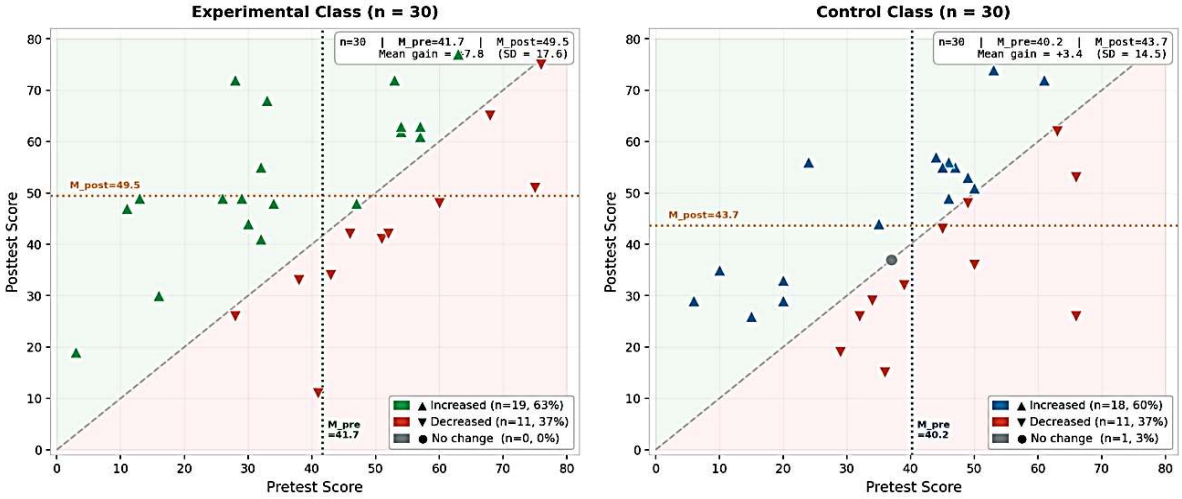

Figure 5. Individual pretest–posttest score distribution per student

Descriptively, observational data from the Commognitive Component Observation Sheet showed that the experimental class had substantially higher mean scores for Word Use (M = 3.2 vs. 2.4), Visual Mediators (M = 3.1 vs. 2.3), Routines (M = 3.0 vs. 2.5), and Endorsed Narratives (M = 2.9 vs. 2.1) compared to the control class. Similarly, the Discussion Role Change Rubric revealed a higher frequency of Initiator–Elaborator–Evaluator transitions in the experimental class (mean transitions per session: 4.8 vs. 1.9). These patterns suggest that greater activation of commognitive components and more dynamic role changes in the experimental class were associated with the higher posttest gains observed. Sequential pattern analysis of the observational data revealed that the most common role transition sequence in the experimental class was Initiator '! Elaborator (occurring in 68% of observed transitions),

followed by Elaborator '! Evaluator (54%), and Initiator '! Evaluator (41%). Notably, Evaluator '! Initiator transitions, indicating a dialectical cycle of critique and re-proposal, were observed in 29% of transitions in the experimental class but in only 8% in the control class. This pattern suggests that students who critiqued a peer's reasoning (Evaluator) were subsequently prompted to propose revised or alternative approaches (Initiator), creating a self-reinforcing cycle of discursive engagement. In the control class, role transitions remained predominantly in the Initiator category with minimal Elaborator or Evaluator involvement, indicating a stagnant discussion pattern in which one or two students dominated without meaningful peer challenge. A transition diagram illustrating these sequential role patterns is recommended for inclusion in future publications and can be constructed from the observational coding data available from the corresponding author. These descriptive patterns are consistent with a mediation hypothesis, whereby commognitive component activation may mediate the effect of the integrative treatment on mathematical problem-solving outcomes: the treatment (integration of Polya's stages with commognitive discourse) appeared to activate commognitive components, which in turn were associated with dynamic role changes, and ultimately with improved problem-solving scores. Future research employing formal mediation analysis, for example, via SmartPLS path modeling, is recommended to rigorously test this indirect pathway (treatment '! commognitive activation '! role change '! problem-solving ability) and to quantify the relative contribution of each mediating variable.

Paired T-Test Results

Separate paired samples t-tests were conducted for each class. For the experimental class, the paired t-test yielded $t(29) = 2.435$, $p = 0.021$, indicating a statistically significant improvement from pretest ($M = 41.67$, $SD = 18.65$) to posttest ($M = 49.50$, $SD = 16.45$). For the control class, the paired t-test yielded $t(29) = 1.295$, $p = 0.206$, indicating no statistically significant improvement from pretest ($M = 40.23$, $SD = 15.64$) to posttest ($M = 43.67$, $SD = 15.28$). These results demonstrate that the integrative Polya–commognitive treatment produced a statistically significant within-class improvement in the experimental group ($p = 0.021$). In contrast, the control group did not show significant pre-post gains ($p = 0.206$). To compare the effectiveness of the two conditions, an independent samples t-test was conducted on N-Gain scores. The experimental class achieved a mean N-Gain of 0.08 ($SD = 0.32$, low category), compared with the control class's mean N-Gain of 0.03 ($SD = 0.30$, low category). The independent t-test on N-Gain yielded $t(58) = 0.699$, $p = 0.487$, indicating no statistically significant difference in normalized gain between the two classes. Nevertheless, the significant within-group improvement in the experimental class, in contrast to the non-significant result in the control class, provides evidence that the structured integrative treatment was associated with meaningful learning gains, even as the overall effect magnitude remained modest. This contrast reflects the distinction between the statistical significance of within-group change and that of between-group comparisons, both of which are necessary for a complete evaluation of treatment effectiveness, active, reflective, and argumentative roles. Thus, the results of this test provide strong empirical evidence that the interaction between problem-solving processes and mathematical discussion activities significantly improves the quality of thinking and learning outcomes.

Relationship between Polya's Stages and Commognitive Components

The improvement in posttest scores is consistent with theoretical expectations that

Table 10. Paired samples t-test

			statistic	df	P
Experimental Class	Pre → Post	Student's t	2.435	29	0.021
Note. H_a $\mu_{Measure\ 1 - Measure\ 2} \neq 0$					
Note. H_a $\mu_{Measure\ 1 - Measure\ 2} \neq 0$					

Table 11. Independent samples t-test on n-gain scores

Class	N	Mean N-Gain	SD	t (df=58)	p
Experimental	30	0.084	0.323	0.699	0.487
Control	30	0.027	0.301		

Note. H_a: $\mu_1 \neq \mu_2$. t and p values correspond to the independent samples t-test comparing N-Gain scores between the two classes.

integrating heuristic stages with discursive practices deepens mathematical understanding (Fabiani Marcatto, 2025; Oktaviani et al., 2020; Wiyah & Nurjanah, 2021). Observational data showed that higher commognitive activation, especially Word Use and Endorsed Narratives, was concentrated during the Planning and Evaluation phases (Polya P2 and P4). Students who consistently used precise mathematical terminology during P2 discussions tended to score higher on P4 evaluation items, suggesting a meaningful within-student association between discursive quality and reflective problem-solving competence. This aligns with findings by Martín Molina et al. (2020) and Tyskerud et al. (2023), who identified that commognitive conflicts at the planning stage promote definitional clarification and strategy refinement. When managed productively, cognitive conflicts can also lead to the correction of strategies, thereby improving Polya's planning and evaluation stages (Gavilán Izquierdo et al., 2022). Furthermore, studies on commognitive discourse specifically highlight that limited use of mathematical vocabulary and weak justification narratives, as observed in the control class, are associated with lower-quality problem-solving outcomes (Sihlangu et al., 2025). The commognitive mechanism provides a conceptual explanation for why explicit discourse-oriented treatment improved performance beyond procedural rehearsal alone.

Relationship between Commognitive Components and Role Change Dynamics

Observational data indicated that higher-quality activation of the commognitive component coincided with more frequent and varied role transitions in the experimental class. Specifically, sessions with high Endorsed Narrative scores were associated with more Initiator-to-Evaluator role transitions, as students who constructed explicit justifications were more likely to challenge and evaluate peers' reasoning. This pattern supports the theoretical claim by Ben-Dor & Heyd-Metzuyanim (2021) that meta-level learning agreements, where students negotiate mathematical objects discursively, are associated with shifts in dynamic positioning. The distribution of epistemic authority through dynamic role changes has also been identified as a key determinant of the quality of meaningful collaborative learning (Tran & Díez-Palomar, 2021). In contrast, the control class showed predominantly Initiator roles with minimal Elaborator or Evaluator transitions, consistent with the stagnant discussion patterns reported by Fatmanissa et al. (2022) and Sutama et al. (2022). The observational data thus provide direct

empirical support for the association between commognitive activation and role-change dynamics.

Relationship between Polya's Stages and Role Change through Discursive Practices

The link between structured problem-solving stages and role-change dynamics was most evident during the Evaluation and Reflection phase (P4). In the experimental class, the structured requirement for each group to present and defend its solutions publicly created conditions for epistemic role shifts: presenters assumed the Initiator role; peers assumed the Elaborator or Evaluator role. This structured format mirrors the findings of Muslim et al. (2024), who documented that Polya's evaluation stage, when supported by commognitive discourse, generates the highest frequency of meaningful role transitions. The lecturer's facilitative questioning—prompting justification rather than providing answers—was crucial in sustaining exploratory discussion, consistent with evidence from Gustafsson (2024) and Liu & Cao (2025). Furthermore, teacher or lecturer intervention in facilitating discussions has proven crucial for keeping them exploratory and preventing them from falling into a mechanistic division of labor (Guo et al., 2025; Liu et al., 2025).

Measurement Quality and Methodological Reflections

Rasch evidence confirmed that instrument scores reliably reflect variation in ability rather than measurement artifacts, thereby strengthening the validity of observed score differences between groups. The parallelism of the person–item map indicates that the range of item difficulty is relevant to the population, so that changes in pretest–posttest scores are more credible as representations of ability change, rather than instrument artifacts (Barnett, 2022; Clark et al., 2022; Lefrida et al., 2023). This interpretive robustness is further reinforced when the category structure of the rating scale functions optimally, ensuring that score improvements reflect genuine competency gains rather than ambiguity in the rubric (Weingarden & Heyd-Metzuyanim, 2025). However, several methodological limitations should be noted. The study used purposive sampling within a single institution, limiting generalizability. It is important to address a seemingly contradictory finding in the present results: the paired t-test yielded a statistically significant difference ($p = 0.003$), yet the N-Gain score for the experimental class was only 0.16, falling in the low category. At first glance, these two indicators appear to be in tension. However, they measure different aspects of the treatment effect. Statistical significance (p-value) indicates the probability that the observed difference occurred by chance; with a relatively controlled quasi-experimental design and an adequate sample size, even small real differences can yield significant p-values. The N-Gain, by contrast, measures the magnitude of improvement relative to the maximum possible gain, and a value of 0.16 indicates that the absolute gain was modest in proportion to the ceiling of the instrument. A plausible explanation for this discrepancy is the brevity of the treatment period: six two-hour sessions may be insufficient to produce large-scale gains in mathematical problem-solving ability, which typically requires sustained practice over an extended period. The commognitive discourse approach, while theoretically promising, may require more time for students to internalize and apply new epistemic roles fluently. Future studies should consider extending the treatment duration to at least one full semester to allow for deeper consolidation of commognitive practices, which may produce N-Gain values in the medium or high category while maintaining statistical significance. The N-Gain values were in the low-to-medium range, suggesting that while the

treatment effect was statistically significant, its practical magnitude was modest. The structural relationships among the three constructs, Polya's stages, commognitive components, and role change, were examined through observational data analysis and t-tests rather than through structural equation modeling. Furthermore, although the paired t-test shows a significant difference, theoretical readings should focus on why the change occurred. Recent literature emphasizes that modern problem solving requires task designs that create space for explanation, strategy comparison, and reflection, encouraging participants to construct arguments rather than just solve problems (Berman et al., 2024; Johnson & Ohtani, 2025). Technology support or responsive design can also enrich discussions and accelerate meaning negotiations, especially when used to generate alternative solutions and test the plausibility of results (Gustiningsi et al., 2024; Sumalan & Moldoveanu, 2024). In addition, learning practices that foster meta-level learning tend to improve the ability to evaluate and revise strategies, a common weakness in Polya's evaluation stage (Nachlieli & Elbaum-Cohen, 2021; Tseng et al., 2025). Future research is strongly recommended to test these inter-construct relationships using SmartPLS or partial least squares modeling, which would allow simultaneous estimation of direct and mediated paths (e.g., Polya stages '! commognitive components '! role changes '! problem-solving outcomes) and provide a more complete and rigorous test of the theoretical framework proposed in this study.

The consistency of the assumptions of normality (Shapiro–Wilk) and homogeneity of variance (Levene) provides a strong basis for interpreting the differences that emerge parametrically without significant distributional bias. The Shapiro–Wilk test is appropriate for relatively small to medium sample sizes, so the decision to proceed with parametric tests is methodologically justified. Substantively, meeting these statistical prerequisites is important because research on problem-solving discussions often encounters high variability due to differences in participation intensity and group interaction structures. However, when the variances between groups are equal, increases in scores are more likely to reflect the influence of the learning treatment rather than random fluctuations (Gustafsson, 2024; J. Smith, 2024). Thus, the results of the prerequisite test reinforce the validity of the inference that learning integrating Polya's stages and commognitive discourse reinforcement is associated with greater meaningful achievement gains (Muslim et al., 2024; L. Zhang & Ma, 2023).

Overall, the findings support the conclusion that the quality of the connection between Polya's stages and commognitive components, and the dynamics of role changes in discussions, correlate with improved problem-solving achievement, as indicated by stronger posttest differences in the experimental class and statistically significant changes. The theoretical implication is that problem solving needs to be understood as an integrated discursive-cognitive practice: success depends not only on the sequence of steps, but on how terms, visual mediators, routines, and justifications are produced and negotiated within the discussion community (Felmer, 2023; Martín Molina et al., 2020). Follow-up research is recommended to test the model of inter-construct relationships using SmartPLS to map direct and mediating paths, including testing the model's consistency across tasks and class contexts, and considering lecturer intervention as a moderator variable in collaborative discussions.

■ CONCLUSION

This study provides empirical evidence that integrating Polya's problem-solving stages with commognitive discourse practices in collaborative learning is associated with greater

improvement in mathematical problem-solving ability among pre-service teacher students, as indicated by significantly higher posttest scores in the experimental class compared to the control class ($t(29) = 2.435$, $p = 0.021$ (experimental class)). Observational data further suggest that the activation of commognitive components, including consistent use of mathematical terms, visual mediators, structured routines, and justification narratives, and the emergence of dynamic epistemic role transitions during discussion, were associated with improved problem-solving quality in the experimental class. These findings are consistent with the theoretical framework connecting Polya's heuristic stages, commognitive discursive practices, and role change dynamics.

The innovation of this study lies in its attempt to operationalize and measure, within a single quasi-experimental framework, the integration of heuristic stages with commognitive discourse and role-change dynamics. Practically, these findings suggest the value of learning designs that explicitly facilitate exploratory and reflective discussion across all four stages of Polya's framework, rather than focusing solely on procedural execution. Future research should extend this work by employing structural equation modeling to rigorously map direct and mediated inter-construct relationships, explore the moderating role of lecturer intervention and task design, and test the consistency of these findings across diverse institutional contexts and educational levels.

DECLARATION OF GENERATIVE AI USAGE IN THE WRITING PROCESS

During the writing of this manuscript, the author(s) employed Claude AI (Anthropic) to assist with language refinement and proofreading. The author(s) have reviewed and edited the content generated by this tool and assume full responsibility for the content of the published article.

ACKNOWLEDGMENTS

The author would like to express gratitude to the Directorate General of Higher Education, Research, and Technology through the Master Contract No. 129/C3/DT.05.00/PL/2025 dated May 28, 2025, as well as the Subcontract No. 2166/LL8/AL. 04/2025 and 124/004/STKIP-TSB/KETUA/VI/2025 dated June 5, 2025, and June 10, 2025.

REFERENCES

Barnett, J. H. (2022). Primary source projects as textbook replacements: a commognitive analysis. *ZDM - International Journal on Mathematics Education*, *54*(7), 1569–1582. https://doi.org/10.1007/s11858-022-01401-2

Ben-Dor, N., & Heyd-Metzuyanim, E. (2021). Standing on each other's shoulders: A case of coalescence between geometric discourses in peer interaction. *Journal of Mathematical Behavior*, *64*. https://doi.org/10.1016/j.jmathb.2021.100900

Ben-Dor, N., & Heyd-Metzuyanim, E. (2025). Shifts in meta-level learning agreement during mathematics peer learning: integrating positioning theory and commognitive framework. *Educational Studies in Mathematics*. https://doi.org/10.1007/s10649-025-10433-w

Berman, A., Mahagna, A., Ram, I., & Wolf, A. (2024). Problem-solving before instruction: A case study of a matrix theory course. *PRIMUS*. https://doi.org/10.1080/10511970.2024.2352370

Clark, K. M., Can, C., Barnett, J. H., Watford, M., & Rubis, O. M. (2022). Tales of research initiatives on university-level mathematics and primary historical sources. *ZDM - International Journal on Mathematics Education*, *54*(7), 1507–1520. https://doi.org/10.1007/s11858-022-01382-2

Edwards, A. (2025). Trusted together: A commognitive perspective on a primary source project in multivariable calculus. *Mathematics Enthusiast*, *22*(1–2), 123–148. https://doi.org/10.54870/1551-3440.1655

Fabiani Marcatto, F. S. (2025). The three decades journey of problem solving in Brazil. *ZDM - International Journal on Mathematics Education*, *57*(7), 1371–1381. https://doi.org/10.1007/s11858-025-01750-8

Fatmanissa, N., Siswono, T., Lukito, A., & Rahaju, E. B. (2022). Collaborative problem solving in mathematics: A systematic literature review. *Pedagogika*, *148*(4), 45–65. https://doi.org/10.15823/p.2022.148.3

Felmer, P. (2023). Collaborative problem-solving in mathematics. *Current Opinion in Behavioral Sciences*, *52*. https://doi.org/10.1016/j.cobeha.2023.101296

Gavilán Izquierdo, J. M., Gallego-Sánchez, I., González, A., & Puertas, M. L. (2022). A new tool for the teaching of graph theory: identification of commognitive conflicts. *Mathematics Teaching-Research Journal*, *14*(2), 186–212.

Guo, F. Y., Liu, S. Y., Zheng, T. C., & Cao, T. W. (2025). Session - based recommen dation with quaternion-enhanced attention calculation. *Intelligent Data Analysis*, *29*(4), 850–865. https://doi.org/10.1177/1088467X241301378

Gustafsson, P. (2024). Productive mathematical whole-class discussions: A mixed-method approach exploring the potential of multiple-choice tasks supported by a classroom response system. *International Journal of Science and Mathematics Education*, *22*(4), 861–884. https://doi.org/10.1007/s10763-023-10402-w

Gustiningsi, T., Putri, R. I., Zulkardi, Z., & Hapizah, H. (2024). Supporting students' mathematical literacy skill using digital tools. In N. Happy, D. Wulandari, Muhtarom, Sutrisno, M. S. Zuhri, F. Nursyahidah, & D. Purwosetiyono (Eds.), *AIP Conference Proceedings* (Vol. 3046, Issue 1). American Institute of Physics Inc. https://doi.org/10.1063/5.0194695

Iilonga, H. K., & Ogbonnaya, U. I. (2023). Grade 10 namibian learners' problem-solving skill in algebraic word problems. *Infinity Journal*, *12*(2), 275–290. https://doi.org/10.22460/infinity.v12i2.p275-290

Johnson, H. L., & Ohtani, M. (2025). Advances in task design in mathematics education. *ZDM - International Journal on Mathematics Education*, *57*(4), 651–664. https://doi.org/10.1007/s11858-025-01690-3

Lathifaturrahmah, L., Nusantara, T., Subanji, S., & Muksar, M. (2024). Analysis of mathematics students' problem-solving skills in making prediction mathematical representations. In L. Anwar, D. Rahmadani, D. E. Cahyani, T. Listiawan, I. Rofiki, P. Darmawan, & A. D. Pahrany (Eds.), *AIP Conference Proceedings* (Vol. 3049, Issue 1). American Institute of Physics Inc. https://doi.org/10.1063/5.0195480

Lefrida, R., Siswono, T. Y. E., & Lukito, A. (2023). Development of derivative understanding task instruments to explore student commognition. *Mathematics Education Journal*, *17*(3), 343–360. https://doi.org/10.22342/jpm.17.3.20826.343-360

Liu, Q., Wu, J., & Cao, Y. (2025). Deep learning-based low-complexity hybrid precoding for massive MIMO systems. *IEEE Communications Letters*, *29*(10), 2238–2242. https://doi.org/10.1109/LCOMM.

2025.3591774

Liu, Y., & Cao, Y. (2025). Teacher intervention during collaborative problem solving in mathematics classrooms in Mainland China. *Behavioral Sciences*, *15*(3). https://doi.org/10.3390/bs15030377

Lu, J., Wu, S., Wang, Y., & Zhang, Y. (2023). Visualizing the commognitive processes of collaborative problem solving in mathematics classrooms. *Asia-Pacific Education Researcher*, *32*(5), 615–628. https://doi.org/10.1007/s40299-022-00681-2

Martín Molina, V., González-Regaña, A. J., Toscano, R., & Gavilán Izquierdo, J. M. (2020). Differences between how undergraduate students define geometric solids and what their lecturers expect from them through the lens of the theory of commognition. *Eurasia Journal of Mathematics, Science and Technology Education*, *16*(12), 1–10. https://doi.org/10.29333/ejmste/9159

Moustapha-Corrêa, B., Bernardes, A., Giraldo, V., Biza, I., & Nardi, E. (2021). Problematizing mathematics and its pedagogy through teacher engagement with history-focused and classroom situation-specific tasks. *Journal of Mathematical Behavior*, *61*. https://doi.org/10.1016/j.jmathb.2021.100840

Muslim, M., Nusantara, T., Sudirman, S., & Irawati, S. (2024). The causes of changes in student positioning in group discussions using Polya's problem-solving and commognitive approaches. *Eurasia Journal of Mathematics, Science and Technology Education*, *20*(9), em2506. https://doi.org/10.29333/ejmste/15148

Nachlieli, T., & Elbaum-Cohen, A. (2021). Teaching practices aimed at promoting meta-level learning: The case of complex numbers. *Journal of Mathematical Behavior*, *62*. https://doi.org/10.1016/j.jmathb.2021.100872

Oktaviani, N. K. C., Prastika, A. P. A., Fajaroh, F., & Suharti, S. (2020). Incorporation of Polya's problem solving into process guide inquiry in learning buffer solution. In H. Habiddin, S. Majid, I. Suhadi, N. Farida, & I. W. Dasna (Eds.), *AIP Conference Proceedings* (Vol. 2215). American Institute of Physics Inc. subs@aip.org. https://doi.org/10.1063/5.0000692

Sihlangu, S., Maphutha, K., & Mokwana, L. (2025). A commognitive analysis of learners' mathematical thinking on mathematics vocabulary used during classroom discourse. *Journal on Mathematics Education*, *16*(4), 1389–1406. https://doi.org/10.22342/jme.v16i4.pp1389-1406

Smith, J. (2024). Supporting metacognitive talk during collaborative problem solving: A case study in Scottish primary school mathematics. *Education 3-13*, *52*(8), 1578–1593. https://doi.org/10.1080/03004279.2023.2187670

Sumalan, A.-L., & Moldoveanu, C.-E. (2024). Use of digital technology in integrated mathematics education. *Applied System Innovation*, *7*(4). https://doi.org/10.3390/asi7040066

Suparatulatorn, R., Jun-On, N., Hong, Y.-Y., Intaros, P., & Suwannaut, S. (2023). Exploring problem-solving through the intervention of technology and realistic mathematics education in the calculus content course. *Journal on Mathematics Education*, *14*(1), 103–128. https://doi.org/10.22342/JME.V14I1.PP103-128

Sutama, S., Fuadi, D., Narimo, S., Hafida, S. H. N., Novitasari, M., Anif, S., Prayitno, H. J., Sunanih, S., & Adnan, M. (2022). Collaborative mathematics learning

management: Critical thinking skills in problem solving. *International Journal of Evaluation and Research in Education*, *11*(3), 1015–1027. https://doi.org/10.11591/ijere.v11i3.22193

Tran, D., & Díez-Palomar, J. (2021). Measuring agency in mathematics collaborative problem solving. In M. Inprasitha, N. Changsri, & N. Boonsena (Eds.), *Proceedings of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 153–160). Psychology of Mathematics Education (PME).

Tseng, T.-N., Lee, T.-M., & Lin, J.-Y. (2025). Enhancing inquiry-based learning in human factors engineering with generative AI: A case study in industrial design education. In B. K. Smith & M. Borge (Eds.), *Lecture Notes in Computer Science: Vol. 15808 LNCS* (pp. 402–414). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-031-93746-0_28

Tyskerud, A., Mosvold, R., & Bjuland, R. (2023). Commognitive conflicts in mathematics teachers' pedagogical discourse in lesson study. *Mathematics Teacher Education and Development*, *25*(1), 81–92.

Weingarden, M., & Heyd-Metzuyanim, E. (2023). What can the realization tree assessment tool reveal about explorative classroom discussions? *Journal for Research in Mathematics Education*, *54*(2), 97–117. https://doi.org/10.5951/jresematheduc-2020-0084

Weingarden, M., & Heyd-Metzuyanim, E. (2025). Fostering pre-service teachers' attention to mathematical objects: The realization tree mediator as a teaching representation. *Journal of Mathematics Teacher Education*, *28*(4), 851–877. https://doi.org/10.1007/s10857-024-09622-w

Wiyah, R. B., & Nurjanah, N. (2021). Error analysis in solving the linear equation system of three variables using Polya's problem-solving steps. In R. Oktavia, E. Yusibani, Mailizar, Suhartono, Rahmi, Elizar, Irwandi, E. Etkina, G. Planinsic, N. Mansour, N. Idris, C. O'Donnell, K. Kagawa, R. Sheffield, V. M. Mistades, & O. Kaosaiyaporn (Eds.), *Journal of Physics: Conference Series* (Vol. 1882, Issue 1). IOP Publishing Ltd. https://doi.org/10.1088/1742-6596/1882/1/012084

Zhang, L., & Ma, Y. (2023). A study of the impact of project-based learning on student learning effects: a meta-analysis study. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1202728

Zhang, S., Cao, Y., Chan, M. C. E., & Wan, M. E. V. (2022). A comparison of meaning negotiation during collaborative problem solving in mathematics between students in China and Australia. *ZDM - International Journal on Mathematics Education*, *54*(2), 287–302. https://doi.org/10.1007/s11858-022-01335-9

Zhang, S., Esther Chan, M. C., & Cao, Y. (2021). Examining student participation in collaborative mathematics problem solving through the lens of commognitive framework.. In M. Inprasitha, N. Changsri, & N. Boonsena (Eds.), *Proceedings of the International Group for the Psychology of Mathematics Education* (Vol. 1, p. 197). Psychology of Mathematics Education (PME).