Koh, E., et al. (Eds.) (2026). Proceedings of the 2nd International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education.

Investigating Metacognitive Utterances and Reasoning Quality in a Remote Collaborative Japanese Classroom

Lara MONTEAGUDO TUBAU[a*], Changhao LIANG[b], Yu YAN[a], Yu-Tung CHEN[a] & Hiroaki OGATA[b]

[a] *Graduate School of Informatics, Kyoto University, Japan*
[b] *Academic Center for Computing and Media Studies, Kyoto University, Japan*
lara.monteagudo.84h@st.kyoto-u.ac.jp

Abstract: This study examines how metacognitive utterances (planning, evaluation, and monitoring) interact with reasoning quality, while considering individual proficiency. 14 Japanese high school students participated in a remote class to solve math exercises. Using a social metacognition and evidence-based reasoning framework, we analyzed reasoning units to identify metacognitive utterances and eight levels of reasoning quality, which we compared with individual proficiency profiles. The results showed the following differences: (1) High monitoring activity produced high-quality reasoning units, whereas the planning-dominant group produced no high-quality reasoning units; (2) The groups showed no direct peer correction, which may indicate the use of monitoring questions as correction methods that are not openly acknowledged; (3) The combined expertise of group members with varying levels of individual proficiency was the primary driver of collaborative interactions. The study results show that content proficiency does not automatically lead to better collaboration; rather, it requires educators to teach students metacognitive skills, particularly questioning techniques, to succeed in structured group work environments.

Keywords: Social Metacognition, Socially Shared Regulation of Learning (SSRL), Collaborative Learning, Reasoning Quality

1. Introduction

The rise of generative AI intensifies a well-established educational challenge: shifting emphasis from knowledge acquisition to collaborative problem-solving. When information is so easily accessible, students' ability to think together may matter more than individual proficiency. Yet collaborative learning remains underdeveloped. In Japan, research identifies a persistent gap between policy mandates and classroom practice, where instruction prioritizes teacher-guided procedures and correct answers over sustained peer dialogue (MEXT, 2023; Fukuda et al., 2024). This study examines how students regulate collective thinking during remote mathematical problem-solving. Drawing on socially shared regulation of learning (SSRL) frameworks, we define effective collaboration as requiring groups' joint engagement in planning, monitoring, and evaluation (Järvelä & Hadwin, 2013). These metacognitive utterances make reasoning visible and enable groups to co-construct knowledge rather than merely share answers (Halmo et al., 2022; Iiskala et al., 2011). We asked: (1) How are metacognitive utterances and reasoning quality distributed during the remote problem-solving activity? and (2) What is the relationship between metacognitive utterances and collaborative reasoning quality? This research will inform pedagogy for technology-mediated collaborative learning by better understanding collaboration dynamics.

2. Research Background

2.1 Challenges of collaborative learning in Japan

While Japanese teachers maintain consistent constructivist beliefs across all levels, the implementation of active learning significantly decreases in high school (Fukuda et al., 2024). This pedagogical divergence is primarily attributed to the backwash effect of university entrance examinations. These high-stakes assessments mandate a transition from the collaborative, process-oriented norms of elementary education toward a teacher-centered focus on individual content mastery and factual recall at the secondary level. However, because substantive collaboration is a universal challenge rather than a uniquely Japanese one (Isohätälä et al., 2020), it remains critical to examine how collaborative processes manifest within these institutional constraints and which metacognitive strategies students utilize to negotiate shared understanding.

2.2 From individual to social metacognition

Metacognition, thinking about thinking, encompasses planning, monitoring, and evaluation (Flavell, 1979). In collaborative contexts, these processes extend beyond individuals to become socially shared. In SSRL, metacognition is constructed at the group level rather than by individuals acting alone (Järvelä et al., 2013). Unlike individual metacognition or temporary peer scaffolding, SSRL involves the interdependent construction of regulatory activity, with group members building on one another's contributions (Iiskala et al., 2011), where effective functioning shifts discourse from simple answer-sharing toward visible, co-constructed reasoning (Halmo et al., 2022). This distinction is methodologically important: analyzing collaboration requires examining not only whether metacognitive utterances occur but also how they function within interaction. Students with higher proficiency tend to dominate group reasoning, with peers deferring to their contributions (Barron, 2003). The distribution of proficiency, not merely its presence, shapes metacognition. This study examines how individual proficiency profiles moderate metacognitive utterances.

3. Methods

This study employs a sequential mixed-methods design (Creswell & Plano Clark, 2017) to examine how students share their thinking processes and develop reasoning in remote, supported math collaboration. The analysis applies a social metacognition and evidence-based reasoning framework developed by Halmo et al. (2022) in a Japanese remote problem-solving context.

3.1 Group formation and exercise

We assigned 14 first-year high school students (4 urban, 10 rural) to heterogeneous Zoom rooms using GLOBE, a learning log-based grouping tool. Groups were composed based on self-reported proficiency, and each room included at least one urban student to maintain regional diversity (Liang et al., 2021). Collaboration took place in BookRoll, which provided a shared digital workspace and captured reading/annotation logs to support later analysis (Yoshitake et al., 2020).The 60-minute session proceeded in three stages. First, students worked individually on three problems: P1 (Definition of a Function), P2 (Find Vertex), and P3 (Find Maximum Value). Next, students met in breakout rooms to discuss and resolve P1–P3 together, comparing individual approaches and final answers. Finally, in a second breakout session, groups completed a new P4(Maximum of a Quartic Function via Critical Points) which explicitly prompted students to consult the integrated AI chatbot for hints; the chatbot was available throughout. All collaboration and artifact sharing occurred within BookRoll (shared annotation and persistent access), and students submitted a brief post-activity.

3.2 Data sources and processing

Data comprised (1) Zoom recordings (video, audio, transcript) and (2) BookRoll trace logs (pen stroke) (Yoshitake et al., 2020). Verbal data were transcribed verbatim. Ambiguous utterances due to noise or shared microphones were removed after cross-referencing with video and digital traces to verify speaker identity. We segmented transcripts into on-task (mathematical discussion) and off-task (others) categories. To account for remote interaction rhythms, we defined pauses >5 seconds preceding a speaker change, as in Halmo et al. (2022), or >10 seconds of single-speaker silence as qualifying silences, adapting face-to-face benchmarks to the online environment. One group (Room 3) was excluded because of technological challenges that prevented complete data capture.

3.3 Analytical framework and coding

Phase 1: Individual proficiency profiles. We constructed individual proficiency profiles using a 5-point holistic rubric adapted from Charles, Lester, and O'Daffer (1987). Proficiency is the mean score across three individual problems on a 0–4 scale: 4 = correct answer with shown work and minimal errors; 3 = clear understanding with minor conceptual/calculation errors; 2 = on the right track but significant non-trivial errors; 1 = minimal effort or partial understanding; 0 = no attempt or completely incorrect. This adaptation prioritizes mathematical reasoning over accuracy, so minor procedural slips do not disproportionately penalize students, while still accounting for efficiency under time constraints.

Phase 2: Group interaction description. We segmented transcripts into reasoning units, defined as contiguous dialogue focused on a single idea or problem step; a new unit began at a topic shift or when a reasoning unit concluded. Because the class occurred online, we also treated pauses > 5 s before a speaker change or > 10 s of single-speaker silence as unit boundaries to accommodate remote-interaction rhythms.

Metacognitive utterances were labeled (planning, monitoring, self-evaluation) and then reasoning units label on a 0–7 scale to measure the depth of co-constructed reasoning following Halmo 2022. The scale distinguishes between non-transactive (individual) and transactive (collaborative) reasoning based on logical completeness and correctness: (1 = correct/incomplete, 2 = correct/complete, 3 = correct/complete descriptive); and 4–7 indicate transactive reasoning (4 = incorrect, 5 = correct/incomplete, 6 = correct/partially complete, 7 = correct/complete).One primary coder (Coder A) coded all data. For each room, Coder A checked reliability with a different researcher: Coder B for Room 2, Coder C for Room 1, and Coder D for Room 4. Each pair first coded independently, then compared. Ambiguous turns (e.g., barely audible sentences, audio leakage) were excluded after cross-checking recordings and BookRoll traces. Inter-rater agreement per pair was $\kappa = 0.812$ (Room 1, A-C), $\kappa = 0.807$ (Room 2, A-B), and $\kappa = 0.765$ (Room 4, A-D).

Phase 3: Group interaction analysis. We summarized the distribution of planning, monitoring, and self-evaluation across reasoning units to address RQ1. We then described cross-room patterns by viewing these distributions alongside room-level proficiency profiles to address RQ2; given the small sample, we conducted descriptive comparisons only. We retained only mutually agreed labels and excluded ambiguous turns after cross-checking the recordings and BookRoll traces.

4. Results

4.1 Profiles of Metacognitive Utterances, Proficiency, and Reasoning Quality

Because individual proficiency (0–4) and group reasoning quality (0–7) are not directly comparable, each was dichotomized into desirable zones: individual scores 3–4 (somewhat correct) and group scores 4–7 (transactive interaction). Regarding the analysis of individual proficiency (Table 1), distinct patterns emerged across the three environments. In Room 1, the mean individual score was 2.50, with half of all scores reaching the desirable threshold (3–4), indicating moderate correctness. Room 2 showed the lowest individual proficiency, and half of scores desirable, yet the distribution was most uneven, including a score of zero. Room

3 demonstrated the highest individual proficiency and approximately three-quarters of scores desirable, reflecting consistent problem-solving accuracy.

Table 1: Individual Proficiency Profiles, and Group Reasoning Quality

	Room 1				Room 2				Room 3			
Individual problem	**IP1**	**IP2**	**IP3**		**IP1**	**IP2**	**IP3**		**IP1**	**IP2**	**IP3**	
S1	4	3	4		3	2	0		3	4	1	
S2	0	2	1		1	4	3		4	3	4	
S3	3	2	4		4	0	0		3	4	1	
S4	2	4	1		4	4	1					
Mean and st.dev.	2.5 ± 1.4				2.2 ± 1.7				3.0 ± 1.2			
Group problem	**GP1**	**GP2**	**GP3**	**GP4**	**GP1**	**GP2**	**GP3**	**GP4**	**GP1**	**GP2**	**GP3**	**GP4**
Level per reasoning unit	3	7	1	3	6/7	4	0	7	3/5	5	0	3
Mean and st.dev.	3.5 ± 0.58				4.8 ± 3.0				3.2 ± 2.0			

Table 2: Individual Proficiency Profiles, and Group Reasoning Quality

	Room 1		Room 2		Room 4	
Metacognitive utterance	**Total**	**Relative**	**Total**	**Relative**	**Total**	**Relative**
Planning						
Planning	15	44%	9	16%	23	66%
Monitoring						
Statement to monitor understanding	5	15%	4	7%	2	6%
Correction of another student	0	0%	0	0%	0	0%
Questions to monitor understanding	3	9%	25	45%	3	9%
Request for information	2	6%	5	9%	0	0%
Evaluation						
Evaluation of self	7	21%	11	20%	7	20%
Evaluation of other	2	6%	1	2%	0	0%
Total	34		55		35	

Highlighted in green is the highest value, highlighted in purple the second highest value.

Regarding the analysis of group reasoning units (Table 1), distinct transactivity patterns emerged across the three environments. In Room 1, the mean reasoning quality was 3.50, but only one of four units reached the desirable transactive threshold (4–7), indicating limited genuine cooperation. Room 2 showed the highest reasoning quality, with a mean of 4.80 and three of four units transactive, reflecting sustained collective interaction. Room 3 demonstrated the lowest reasoning quality, with a mean of 3.20 and only one of four units transactive, despite high individual scores. Across all rooms, reasoning quality was not aligned with individual proficiency levels, suggesting that other factors – such as metacognitive regulation – mediated collective success.

The analysis of metacognitive utterances (Table 2) reveals distinct regulatory patterns across the three environments. In Room 1, the profile is characterized by a dual focus on planning (44%) and self-evaluation (21%). However critical interactions were absent, with minimal evaluation of others (6%) and a total lack of peer correction. Room 2 had the highest frequency of metacognitive activity (n = 55), dominated by questions to monitor understanding (45%) and self-evaluation (20%). Despite this, the group still lacked critical interaction, failing to provide corrective feedback or oppositional claims. Room 4 demonstrated the least balanced distribution, with a focus on planning (66%), suggesting a procedural orientation. Across all groups planning utterances always present while monitoring utterances varied greatly. None of the groups showed significant critical behavior towards other students, such as evaluation of others and correction of another student.

4.2 Associations Between Metacognitive Profiles and Collaborative Reasoning

At the room level, the relationship between metacognitive utterances and collaborative reasoning quality was clearly differentiated across groups. Room 2 combined the highest

metacognitive activity (55 utterances) with highest aggregated monitoring utterances (61%) and the highest mean in unit reasoning quality (4.8, above threshold) (see Table 1). In contrast, Room 4 showed 35 utterances with planning = 66% and the lowest mean reasoning unit quality (3.2, below threshold). Meanwhile, Room 1 fell between these cases in both metacognitive profile and reasoning outcomes (see Table 1 and 2).

Regarding proficiency alongside activity, Room 4 had the highest average proficiency (3.0) yet produced the fewest metacognitive utterances (35) with a planning-heavy pattern (66%). By contrast, despite having lower individual scores (mean 2.17), Room 2 produced more utterances (55) and a monitoring-dominant pattern (45%) than Room 1 (mean 2.50, 36 utterances). Across rooms, corrections and group evaluations were rare or absent.

Finally, Room 4 combined the highest average proficiency (3.0) with the lowest mean group reasoning quality (3.2, below threshold), whereas Room 2 included students with lower individual scores (mean 2.17) yet attained the highest mean reasoning quality (4.8, above threshold), further illustrating how metacognitive profile, not proficiency alone, related to reasoning quality. This is consistent with previous findings. Comparing the rooms, variations in metacognitive utterances (type and distribution) and individual proficiency levels together shaped the differences in reasoning outcomes.

5. Discussion and limitations

5.1 Interpretation of Key Findings

Across groups, the type of metacognitive utterances aligned more strongly with reasoning quality than the overall number of metacognitive utterances. In particular, in Room 2 we observed Questions to monitor understanding (45%) co-occurred with transactive-zone reasoning quality (mean= 4.8), whereas planning-heavy profiles, as in Room 4 (66), corresponded with non-transactive reasoning quality (3.2). Evaluation of self utterances occurred at similar rates across rooms (19–20%) and did not differentiate outcomes.

Monitoring questions likely helped make reasoning visible and negotiable, prompting clarification, elaboration, and mutual checking that support shared understanding. This interpretation aligns with research on social metacognition and co-construction in small-group problem solving, where monitoring functions as a catalyst for transactive moves rather than mere answer-sharing (Halmo et al., 2022; Iiskala et al., 2011).

Group reasoning quality did not simply sum individual proficiency. Room 4 had the highest average proficiency (3.0) yet the lowest reasoning quality (3.2, below the transactive threshold), whereas Room 2 contained lower individual scores (2.2) but reached the highest reasoning quality (4.8, above the transactive threshold). Relatedly, groups with high-proficiency members can underperform when interaction patterns suppress active collaboration or invite deference to a few contributors (Barron, 2003).

Groups with somewhat similar mean proficiency nonetheless developed different metacognitive profiles (Room 1 vs. Room 2), indicating that group composition alone does not determine collaborative regulation. Across rooms, peer corrections and group evaluations were rare or absent, which may reflect cultural or developmental constraints or peer-norm sensitivities in this context. These observations point to the potential value of explicit metacognitive training (e.g., modeling and practicing monitoring questions as socially acceptable forms of “checking” one another’s thinking)

In technology-mediated, small-group mathematics, it may be more productive to train and scaffold specific metacognitive utterances, especially monitoring, than to assume that high individual proficiency will result high-quality collaboration. This implication resonates with broader SSRL perspectives emphasizing joint regulation and interactional construction of planning, monitoring, and evaluation during problem solving (Järvelä et al., 2013; Iiskala et al., 2011; Halmo et al., 2022).

5.2 Limitations and Implications

This exploratory study is limited by a small sample size (12 students, 3 rooms), a one-off design that cannot capture the development of group norms over time, and reliance on consensus coded data, which may undercount certain categories. Technical constraints in the online environment may also have affected participation. We also did not examine individual factors beyond proficiency (e.g., personality, prior acquaintance), log data of tool use, or how the math problem's single correct answer may have suppressed contributions. Analyses remain exploratory, with no statistical testing and inherent coding bias. These factors constrain generalizability and warrant cautious interpretation.

Future work should integrate objective traces, log data from the Learning Management System log data, and chatbot logs to examine how external resources shape metacognitive utterances and reasoning (Yoshitake et al., 2020). Learning-log analytics could support automated group formation and longitudinal study of interactional trajectories (Liang et al., 2021). More granular analysis of within-group proficiency variation and personality factors may clarify when heterogeneous grouping encourages balanced, transactive reasoning and how roles moderate utterances.

Acknowledgements

This work is partly supported by JSPS KAKENHI Grant Number 25K21357 and JST CSTI SIP Program Grant Number JPJ012347.

References

Barron, B. (2003). When smart groups fail. The Journal of the Learning Sciences, 12(3), 307–359. https://doi.org/10.1207/S15327809JLS1203_1

Creswell, J. W., & Plano Clark, V. L. (2017). Designing and conducting mixed methods research (3rd ed.). SAGE Publications.

Charles, R., Lester, F., & O'Daffer, P. (1987). How to evaluate progress in problem solving. Reston, VA: National Council of Teachers of Mathematics.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. American Psychologist, 34(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Halmo, S. M., Brem, S. K., & Beghetto, R. A. (2022). “Oh, that makes sense!”: Social metacognition in small-group problem solving. Instructional Science, 50(1), 105–135. https://doi.org/10.1007/s11251-021-09569-y

Iiskala, T., Vauras, M., Lehtinen, E., & Salonen, P. (2011). Socially shared metacognition of dyads of pupils in collaborative mathematical problem-solving processes. Learning and Instruction, 21(3), 379–393. https://doi.org/10.1016/j.learninstruc.2010.05.002

Isohätälä, J., Näykki, P., & Järvelä, S. (2020). Convergences of joint, positive interactions and regulation in collaborative learning. Small Group Research, 51(2), 229–264. https://doi.org/10.1177/1046496419867760

Järvelä, S., & Hadwin, A. F. (2013). New frontiers: Regulating learning in CSCL. Educational Psychologist, 48(1), 25–39. https://doi.org/10.1080/00461520.2012.748006

Liang, C., Majumdar, R., & Ogata, H. (2021). Learning log-based automatic group formation: System design and classroom implementation study. Research and Practice in Technology Enhanced Learning, 16(1), 14. https://doi.org/10.1186/s41039-021-00156-w

MEXT (Ministry of Education, Culture, Sports, Science and Technology). (2023, June 16). Basic plan for the promotion of education (Cabinet Decision). https://www.mext.go.jp/content/20240228-soseisk02-100000597_09.pdf

Mitsugi, M., Hiromori, T., Yoshimura, M., & Kirimura, R. (2024). The influence of emergent and assigned leaders on interactive group work tasks in the L2 classroom: Focusing on the group work dynamics, motivation, and linguistic performance. The Journal for the Psychology of Language Learning, 6(1), 1–21. https://www.jpll.org/index.php/journal/article/view/156

Fukuda, M., Fukaya, T., & Kusumi, T. (2024). Differences and Relationships Between Teachers’ Pedagogical Beliefs and Teaching Strategies Used at Different School Levels in Japan. Sage Open, 14(3). https://doi.org/10.1177/21582440241281852

Yoshitake, D., Flanagan, B., & Ogata, H. (2020, November). Supporting group learning using pen stroke data analytics. In International Conference on Computers in Education (pp. 634–639). http://hdl.handle.net/2433/259798