Examining the Relationship Between Calibration and Reflection in an Online Discussion Environment

Yu Xia and Marcela Borge
yzx64@psu.edu, mbs15@psu.edu
The Pennsylvania State University

Abstract: Calibration plays a critical role for groups to regulate future learning behaviors. There is a wealth of research on self-assessment, calibration, and metacognitive learning (e.g., Azevedo, 2009; Zimmerman, 2002), but there is a lack of research on calibration at the group level. To accurately calibrate, an accurate understanding of the assessment criteria plays an important role, and reflective activities in groups on self-assessment could provide opportunities for learners to take time thinking about their performance and correcting each other's understanding of the criteria. This study aims to examine the relationship between learner calibration, at the group and individual levels, with their patterns of discourse in the self-reflective discussions that indicated learner understanding of the assessed items. Results show that high calibration accuracy groups had more accurate understanding of the items, while incorrect understanding in medium and low calibration accuracy groups were either ignored or agreed upon, rather than challenged.

Keywords: calibration, group reflection, metacognition, collaborative learning

Introduction

Collaborative competence is becoming increasingly important in our globalized society where many complex problems require multidisciplinary teams to devise solutions. As necessary collaborative learning is, working collaboratively in groups can introduce new problems that can negatively affect learning outcomes. These problems can be caused by individuals or by interactions between individuals, which is why many argue that high quality collaboration requires metacognition at different levels of scale (Borge, Ong, & Rosé, 2018; Järvelä & Hadwin, 2013). At individual level, learners with better self-regulated learning skills perform better academically (Azevedo, 2009; Zimmerman, 2002); at the group level, groups with higher regulation show more productive processes (Järvelä et al., 2015). Moreover, different self-assessment methods have been shown to have positive effect on self-regulation (Panadero, Tapia, & Huertas, 2012). During self-assessment, learners make judgements about their performance and these judgments are critical to their strategic behaviors in the future (Alexander, 2013). However, research shows that people are generally inaccurate when assessing their own performance (Brown & Harris, 2014; Dunning, Heath, & Suls, 2004; Puncochar & Fox, 2004), which is why learners need support to effectively self-assess (Brown & Harris, 2014). This support includes opportunities to learn how to calibrate self-assessments (Hacker, Bol, & Bahbahani, 2018).

Calibration refers to the extent of alignment between learners' estimated perceptions of their own levels of understanding, capability, or competence, and that of their demonstrated levels, which reflects how accurate they can assess their own performance (Alexander, 2013; Hadwin & Webster, 2013; Pieschl, 2009). Self-assessment activities, where learners make judgments about their performance based on understanding of assessment criteria, can be designed to help students calibrate how they assess themselves, while providing guidance for selecting strategies to help them improve. It can be justifiably hypothesized that self-reflective activities could potentially help learners with their calibration since such activities provide opportunities to adjust their understanding and thus contribute to improving calibration in self-evaluation. Hacker, Bol, and Bahbahani (2008) conducted an experimental design to see the effect of extrinsic incentives and reflection on students' calibration of exam performance, and confirmed previous research findings (Kruger & Dunning, 1999) that high performing students were more accurate in calibration while low performing students were less accurate. However, their findings showed no significant effect from their interventions on calibration. This finding may have been due to a ceiling effect for high performing teams and no effect of the reflection intervention for low performing groups.

Calibration has been a well-researched area for individual learning, but very little research has been conducted at the group level. To address calibration at group level and to improve student socio-metacognitive competence, an online discussion system was designed and developed by Borge et al. (2018). In the present study, we examine how students used the online tool, designed to help teams reflect on and improve collaborative sense-making processes, to individually self-assess their performance and as a group reflect upon the performance and

their self-assessments over one semester. We investigate the relationship between self-assessment accuracy for high, medium, and low calibration teams, at the group and individual level, with their patterns of discourse in the self-reflective discussions that show whether they learned the six core communicative behaviors in collaborative process. Specifically, we ask the following: RQ1) How does the accuracy of group self-assessment scores on collaborative process change over time for high, medium, and low calibration teams? RQ2) How do high, medium, and low calibration teams calibrate their understanding of criteria during the group reflection?

Methods

Research settings

The study took place in a 15-week university level introductory online course in College of Information Sciences and Technology in a university in the Northeast United States. Participants were 27 undergraduate students with the majority being male (70%, n=19; female: 30%, n=8). Twelve teams were formed based on students' availability for synchronous online discussion meetings and their expertise in the course area. Teams carried out five online discussions in weeks four, six, eight, ten, and twelve. Each student completed a Reading Questions Activity where they responded to reading questions and the responses could be referred to in online group discussion. Two groups did not do their self-assessment and justifications and an additional one group did not do their reflection afterwards, which leaves us with nine groups for analyses.

The online text-based discussion environment is called CREATE (*Collaborative Regulation, Enhanced Analysis, and Thinking Environment*). On the home page, there are 6 short videos (less than 6 minutes) explaining each criterion with examples that students can watch before having their discussion. The six videos were designed to help students understand the purpose of each criterion, what meets the criterion, and what an example would look like. When teams meet online, they discuss predetermined questions about course materials for 60-90 minutes, move to individual self-reflection, where they assess discussion quality, and close with a group discussion on individual reflections for the purpose of diagnosing problems and selecting strategies to improve. Students use the same rubric for self-assessment that experts use to assess team discussion quality. The rubric has a 5-point rating scale for six items: verbal equity, developing joint understanding, joint idea building, exploration of different perspectives, quality of claims, and norms of evaluation (for definitions of each item see Borge et al., 2018). The system provides concrete examples of what each score for each item looks like. For example, if students selected. a score of 4 for quality of claims the system would show, "There is at least one example where claims are supported by references to course readings or online content, but no examples of critically analyzing claims" (see Figure 1).

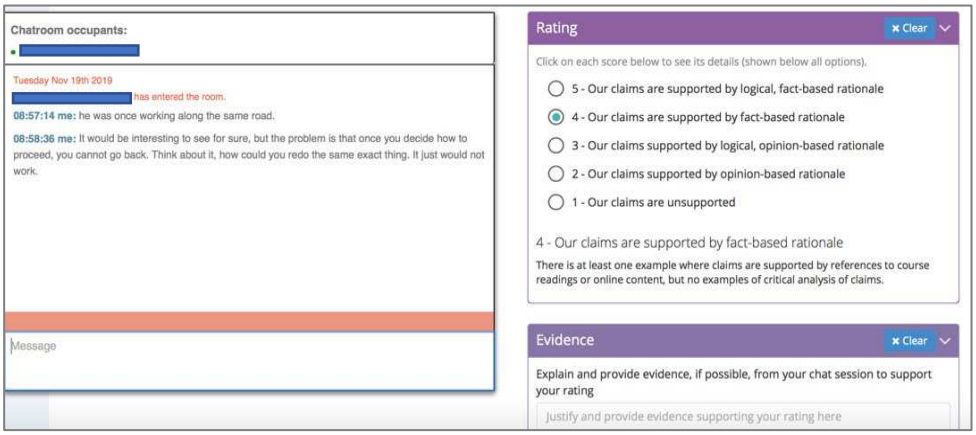

Figure 1. Screenshot of individual reflection feature in the CREATE system.

Data analysis

Calibration. To see whether groups could assess their discussion quality more accurately over time, we examined nine groups' average scores as compared to expert scores for the five discussions. The expert scores were provided by a graduate student after 20% of the data were double coded by two trained graduate students with extensive communication analysis experience and significant agreement was reached between them ($r=.86$, $p<.001$; Kappa=.64, $p< .001$). The expert scored the group performance on the six items, giving one score to one item for each team. Individual member's assessing scores were averaged for group discussion quality. The closer the average score gets to the expert score, the more accurate the group assesses themselves. Overtime, if the difference

between group score and expert score gets closer to 0, it shows higher calibration; if the difference gets further from zero, it shows lower calibration. How well groups calibrated over five discussions, the size and direction of the differences over time, was used to classify groups into the high, medium, and low calibration groups.

Characterizing justification accuracy at the level of the individual. We used content analysis (Chi, 1997) to examine the extent to which individuals understood the items they were asked to assess in the system, items that pushed them to examine collaboration processes. Every time students were asked to score a reflective item, they were also asked to provide evidence in the form of written text to justify the score. We coded students' individual justifications, or rationale, for why they assessed each item as they did. These justifications could contain multiple sentences but were coded at the level of the utterance. The nine groups yielded 898 justification utterances. Utterances were identified and coded based on demonstrated understanding of the rubrics (see Table 1 for code definitions). Two raters coded 18% of responses as justification in their individual reflection phase, for a Cohen's Kappa inter-rater agreement of 0.90. Disagreements were resolved through discussion.

Table 1. Justification accuracy codes

Categories	Description
Match	The rationale for an item score matches up to the definitions/descriptions of that item in the CREATE system.
Mismatch	The rationale for one item does not match the one they are assessing but another item in the CREATE system, which indicates the confusion between assessment criteria and mistaken understanding.
Other	The rationale for an item score does not relate to the descriptions of any item in the CREATE system, or the student does not give any justification, or the rationale is exactly same worded as the description in rubrics.

Characterizing justification accuracy at the level of the group. After students completed their individual reflections, they were to discuss their scores and rationale for scores, collectively, so as to agree on strengths and weaknesses and devise collective plans for improvement. We analyzed these group reflective discussions to examine the relationship between group reflection and individual assessment calibration over time. We first ranked the group calibration accuracy at the last time point, thus forming three high, three medium, and three low calibration groups. A further qualitative analysis was then conducted on all nine groups' group reflection. The same justification accuracy codes (Table 1) were used to assess group justification accuracy. A second round of coding (Table 2) was developed to determine how group members responded to inaccurate justifications, shifting the focus on to what extent they displayed incorrect (Mismatch) or unclear (General) understanding of the items, and more importantly, how their group members reacted to the inaccurate descriptions.

Table 2. Characterizing how others within a team responded to inaccurate justifications

Categories	Description
Ignore	When one student gives an incorrect understanding (Mismatch), or when an unclear or general (General) statement is made, that incorrect or unclear description is ignored.
Agree	When one student gives an incorrect understanding (Mismatch), or when an unclear or general (General) statement is made, other group members show agreement with that incorrect or unclear description.
Challenge	When one student gives an incorrect understanding (Mismatch), or when an unclear or general (General) statement is made, other group members show disagreement with that incorrect or unclear description.

Findings

Changes in group self-assessment accuracy of collaborative process quality

To see the changes in the group accuracy of collaboration process quality, we calculated the differences between expert scores and group average scores by deducting group averages from expert scores. That is, 0 means group average score is the same with expert score, positive results mean the group score is lower than the expert score, thus indicating underestimation, and negative results mean expert score is lower than the group score, indicating overestimation. Based on the analysis result of the group accuracy, we ranked the nine groups, classifying the top three groups as high calibration groups, the next three as the medium calibration groups, the last three as low calibration groups (see Figure 2). In Figure 2, the closer to 0, the more accurate the group was in assessing a particular item at a particular time point. Take Group 3 as an example, the orange line of the item Joint Idea

Building (JIB) shows they were perfectly accurate in assessing this item in sessions 1, 2, and 5, but slightly overestimating their performance at sessions 3 and 4.

Looking at the three high calibration groups (Groups 3, 8, and 12) with the narrowest differences at session five, there is a tendency for narrowing the differences between group average scores and expert scores, and thus becoming more accurate in calibration. The medium calibration groups (Groups 7, 9, and 15) showed improvements as indicated by the narrowing change, but not enough improvements as the high calibration groups. The low calibration groups (Groups 1, 4, and 11) showed no improvements, or even deteriorating accuracy as in the case of Group 1 with wider differences.

As described above, after each group member assessed the group performance, they would get back together as a group to discuss their self-assessments, and then select two items of the total six as their strength and weakness respectively. This was designed to provide an opportunity for groups to check each other's understanding and improve it through talking about each item as a group. However, from the different accuracy results for nine groups over five sessions, there is a need to conduct an exploratory examination into the group reflection discussion to see how they used the learning opportunity to improve their understanding of the collaborative discussion criteria, or not.

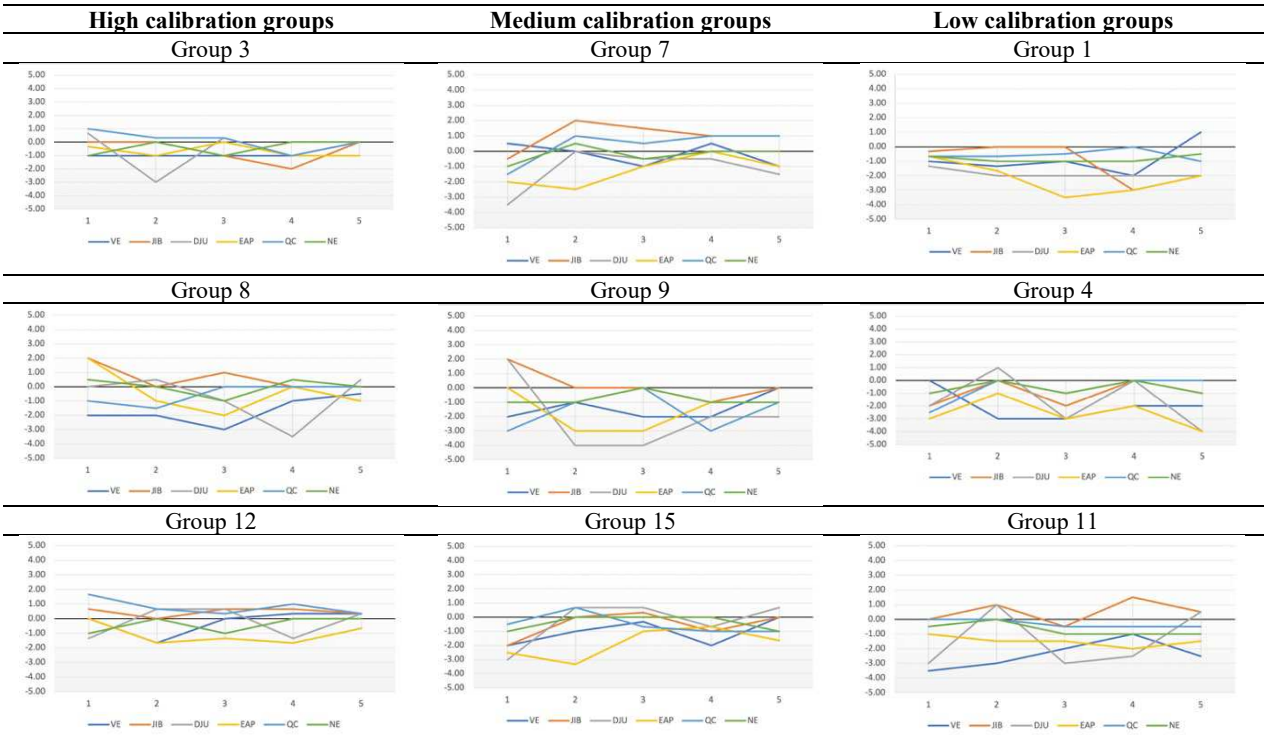

Figure 2. Differences between group average scores and expert scores for each team, from session 1 to 5, and accuracy classification.

Group calibration and group reflection

We coded individual justification entries over five time points for each team to see how well these justifications matched the item description provided by the system as part of the individual self-assessment features. Justifications could be coded as "Match", indicating alignment with specific item descriptions for the score, "Mismatch", indicating confusion between different item definitions or examples, and "Other", indicating no alignment to any item in the system, or "Missing", indicating that the individual did not enter any justifications for their scores. From Figure 3, we can see that, for the three high calibration groups, there were no missing entries, and there were less mismatches than matches. For Group 3, matches (from 53.13% at session 1 to 64.29% at session 5) were much higher than mismatches (from 12.50% at session 1 to 7.14% at session 5); for Group 8, matches (from 50.00% at session 1 to 20.00% at session 5) were slightly more than mismatches (from 0.00% at session 1 to 6.67% at session 5); for Group 12, matches (from 50.00% at session 1 to 61.54% at session 5) were also more than mismatches (from 27.78% at session 1 to 7.69% at session 5). While there was a decrease in the accurate descriptions and even an increase of inaccurate descriptions for Group 8, the overall trend showed that the group had a more accurate understanding of the items. These results indicate that high calibration groups

showed learning of their knowledge of desired communicative behaviors in collaborative discussions, as compared to the rest of the groups.

For medium calibration groups, Group 7 and Group 15 showed an increase in missing entries. Group 7 also had fewer accurate entries over time, and Group 15 remained stable, which shows little learning of the process knowledge of communicative behaviors. While Group 9 showed an increase in accurate understanding, from Table 4 which summarized data from analyzing their group reflections, Group 9 actually had a high number of inaccurate descriptions. For low calibration group, there were fluctuations in both matches and mismatches, but the "Other" category showed overall increase. Group 11 showed increasing in accurate understanding, thus learning of collaborative process knowledge, at sessions 3 and 4, but they regressed at session 5. As these results show, the changes were not linear and consistent. We then further analyzed the reflection discussion and specifically examined how the group reacted when there was an inaccurate or unclear description of any item presented.

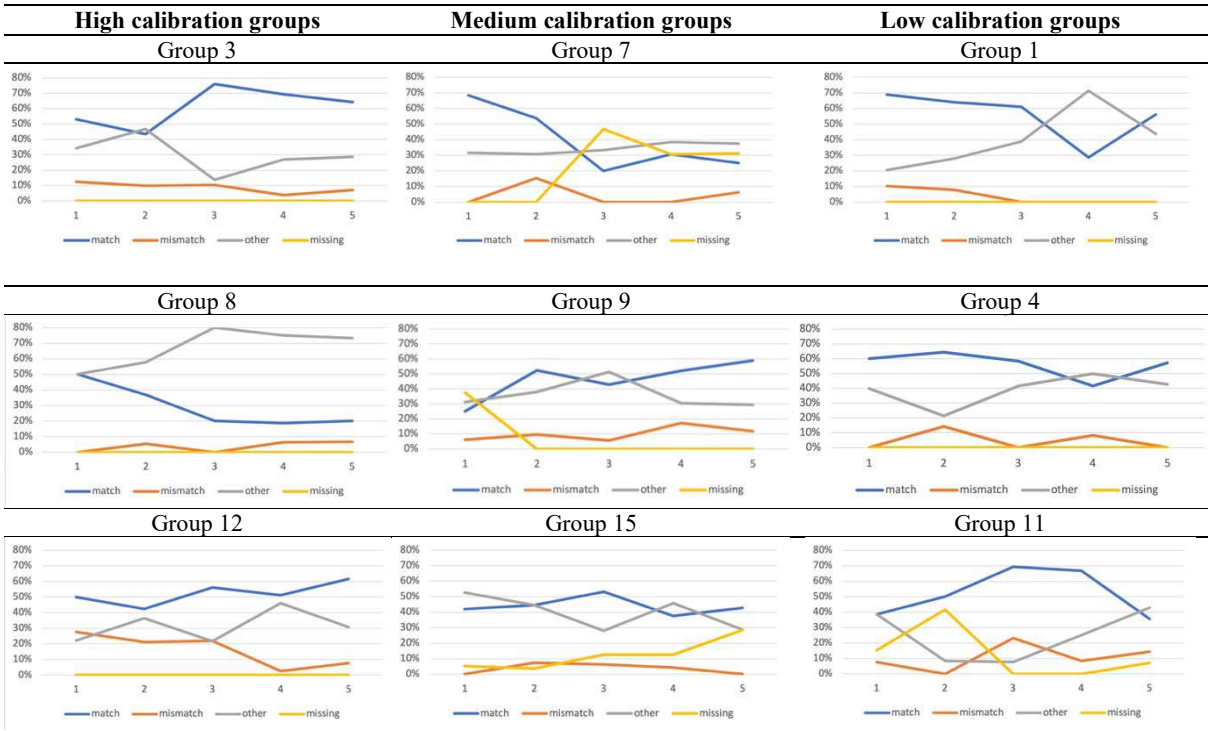

Figure 3. Proportion of changes of individuals' understanding of the self-assessment items within each team.

To further examine the differences of accuracy change across high to low calibration teams, we coded the group reflection discussion, when teams discussed their scores and justifications to collectively diagnose the team and select appropriate strategies to improve their discussion processes. We coded all five discussions over the semester to examine how they justified their scores to the group and see if they engaged in calibration. We coded discussions for the accuracy of the justifications discussed as well as how teammates responded to these justifications. Challenges would indicate calibration. In the excerpt shown in Table 3, Group 9 was discussing their individual scores and justifications for the item "exploration of different perspectives". One person (ID 96) provided justification that was not related to the scoring criteria for alternative perspectives: the number of examples provided. However, no one challenged the incorrect justification, the third person simply agreed with what was being said.

Table 3. Excerpt from Group 9 group reflection

Turn	ID	Post	Code	Note
1	97	*We explore different perspectives and discuss them in depth*	**General**	They were discussing the item of exploring different or alternative perspectives, but this line was a general description.

2	96	*we use a lot of examples for sure*	**Mismatch**	This response did not meet the description of alternative perspectives.
3	96	*and yes different perspectives*	**General**	96 did not specify different perspectives to what, since students easily had a shallow understanding that any differences suffice as an alternative perspective, but the rubric is specifically requiring an alternative perspective on any claim presented by another group member.
4	96	*I'm sure if we had problems we will discuss in depth*	**N/A**	No code applied for this line as it did not speak to any of the items
5	97	*Right*	**Agree**	This is an agreeing with previous responses.

Table 4 shows both the accuracy of their item descriptions as well as how team members responded to item descriptions. There were fair amounts of accurate descriptions across all teams, but there were far less inaccurate descriptions in high accurate groups, with only one instance identified for Group 3. However, of the 23 instances of inaccurate descriptions that occurred in Group 9's collective discussion, none were challenged nor questioned; they were either ignored or agreed upon. Table 4 shows that there is a tendency for group members to discuss the items and their self-assessment in a general way without taking time to rethink those items and without providing evidence to their claims about the items.

Table 4. Group discussion on their self-assessment of collaborative discussion quality

Calibration	Group	Total line	Match	Mismatch	General	Ignore	Agree	Challenge
High	3*	242	13	1	8	4	0	0
	8*	161	5	0	6	1	4	0
	12†	57	6	0	0	0	0	0
Medium	7	68	3	0	3	1	1	0
	9*	345	7	22	13	7	21	0
	15	130	5	0	5	0	0	0
Low	1	227	13	5	4	5	3	0
	4	435	0	10	7	4	10	0
	11‡	58	2	2	0	2	0	0

*Group 3, Group 8, and Group 9 did not have the reflection discussions at session 1.
†Group 12 did not have the reflection discussion at sessions 3 and 4.
‡Group 11 did not have the reflection discussion at sessions 4 and 5.

Discussion

In an effort to investigate the relationship between calibration and group reflection, we ranked the teams using an online collaborative discussion tool according to the accuracy in calibration and investigated their calibration accuracy over time and their patterns of discourse in the groups' self-reflective discussions. We strove to see how the accuracy of group self-assessment scores on collaborative process changed over time for high, medium, and low calibration teams and how high, medium, and low calibration teams calibrate their understanding of criteria during the group reflection.

Similar to Hacker and colleagues (2008), there were differences between groups at different levels of performances. In our analysis, high calibration groups demonstrated more accurate knowledge when providing their rationale for self-assessment scores at the individual level, as well as less inaccurate understanding during the group reflection. This finding suggests that knowledge about desired collaborative processes, i.e., concrete models of competency, helped individuals and teams better monitor and diagnose team problems and that this knowledge of collaborative processes was learned by these teams over time. Medium calibration groups showed an increase in missing entries in individual self-assessment and justification, indicating less desire or cognitive space over time to engage in regulation processes. Low calibration groups showed ups and downs in whether individuals correctly justified items or confused the characteristics of one item for another in the text-based, individual assessments. However, the "Other" category, the category for justifications that were not at all connected to descriptions in the system, showed an overall increasing trend, suggesting that they paid less attention to the descriptions of the rubric items displayed by the system over time, or in other words, did not internalize the models the system provided.

After individual self-assessment and justification, teams were required to discuss their self-assessment scores and justifications to help them calibrate their scores. The total line in Table 4 showed no evidence to suggest

that high calibration groups spent more time collectively reflecting than other groups. This suggests that what matters is not the time teams spend reflecting on their processes, but rather how they are reflecting about their processes. For example, Team 12 showed high accuracy in their calibration, but did not spend much time in checking their understanding during the group level reflection across the five sessions. However, when they discussed their justifications for scores, they shared only accurate knowledge. Low calibration groups have less matches and more mismatches during group talk, indicating an absence of learning collaborative process knowledge. These results suggest that process learning at both individual and group level may be critical for calibration and more accurate assessment. However, considering the student accuracy changes, accurate assessment is not necessary for improvement overall. So, there is a complex relationship that coincides with previous models of regulation (Borge & White, 2016), but these models may be increasingly complex for nested form of cognition because of individual and group level factors. We attempted to unpack this complexity by examining whether incorrect or unclear understanding were further discussed, challenged, or corrected. We found that incorrect responses were either accepted or ignored, but never challenged. One explanation is that misunderstandings of reflection items were shared by the entire group. Another explanation is that group members were conflict avoidant and chose not to talk about differences or ask questions about it.

Taken together, high calibration groups showed more knowledge about the model of collaboration articulated through the reflection items in the system, showed higher ability to calibrate over time, and their self-assessment scores were closer to expert scores; such improvement was not found for medium and low calibration groups. However, we still do not know whether this improved ability to calibrate at individual level contributed to more accurate shared knowledge during group reflection, or whether group reflection contributed to individual members' improved ability to calibrate, or whether it is both. More research is needed to examine the relationships between correct understanding of assessment criteria and the ability to make accurate judgements of their own group performance. Future research could investigate these complex relationships by separating conditions of reflection activities, having groups do assessments at different time points, or assessing baseline calibration ability for comparisons to see how the nested relationships change over time. However, as high calibration groups tend to have more accurate understanding of the items, as evidenced by Groups 3 and 12 in Figure 3, this introduced another layer of complexity. What Figure 3 shows was individual members' understanding in each group when they self-assessed their group performance. That is, that high calibration groups were better able to calibrate could be because their individual members were more accurate in understanding the criteria or critical thinkers who would challenge each other. We believe that calibration at both levels of group and individual members is critical to collaboration learning, but how individual members contribute to the group level understanding, how other members in the group help one individual member to calibrate, or more importantly, what the different patterns are in groups with different level of calibration improvement over time, remains to be further researched. With a better understanding of the complex relationships among calibration, understanding, reflection, and regulation, we would be in better position to provide support for group collaborative learning. For example, further research could examine the effect of real-time prompts in systems when students express incorrect or unclear understanding in their individual reflections or during group discussions but nevertheless get ignored or agreed upon, as this study finds that students tend to do.

Conclusion

In constructive and productive collaborative discussions, group members are able to benefit from alternative perspectives without taking offense and to build on each other's idea, adding new things instead of just agreeing. One way to improve their understanding and performance of these critical aspects in high-quality collaborative discussions, is to improve their metacognitive monitoring and their calibration of their performance. However, as indicated in our findings, providing reflective opportunities for students to calibrate their calibration is not enough. They still demonstrated tendency to agree with each other and not use these opportunities to correct or develop their understanding of collaborative discussion processes. Accurate evaluation of their performance does not guarantee regulation in collaborative process. It is not a linear development from improving process knowledge, increasing accuracy in understanding the communicative behaviors, to using that knowledge to regulate coming-up discussions. While we were only able to analyze one class of students over one semester in this study due to the fact that the nature of analyses was very time-intensive, the complexity in the processes of performing, calibrating, and regulating warrants future studies on more groups to see whether the differences between high and low calibration groups, as we found in the current study, hold; whether in-time support would provide bases on which they make calibration decisions; or whether they will take the opportunities to provide alternative perspectives or disagreements so that they calibrate better as a group.

References

Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction*, *24*, 1-3.
Azevedo, R. (2009). Theoretical, conceptual, methodological, and instructional issues in research on metacognition and self-regulated learning: A discussion. *Metacognition and Learning, 4*, 87-95.
Borge, M., Ong, Y. S., & Rosé, C. P. (2018). Learning to monitor and regulate collective thinking processes. *International Journal of Computer-Supported Collaborative Learning*, *13*(1), 61-92.
Borge, M., & White, B. Y. (2016). Toward the development of socio-metacognitive expertise: An approach to developing collaborative competence. Cognition and Instruction, 34(4), 323–360.
Brown, G. T., & Harris, L. R. (2014). The future of self-assessment in classroom practice: Reframing self-assessment as a core competency. *Frontline Learning Research, 2*(1), 22-30.
Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences*, *6*(3), 271-315.
Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*(3), 69-106.
Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning*, *3*(2), 101-121.
Hadwin, A. F., & Webster, E. A. (2013). Calibration in goal setting: Examining the nature of judgments of confidence. *Learning and Instruction*, *24*, 37-47.
Järvelä, S., & Hadwin, A. F. (2013). New frontiers: Regulating learning in CSCL. *Educational Psychologist, 48*(1), 25-39.
Järvelä, S., Kirschner, P. A., Panadero, E., Malmberg, J., Phielix, C., Jaspers, J., ... & Järvenoja, H. (2015). Enhancing socially shared regulation in collaborative learning groups: designing for CSCL regulation tools. *Educational Technology Research and Development*, *63*(1), 125-142.
Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121-1134.
Panadero, E., Tapia, J. A., & Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and individual differences*, *22*(6), 806-813.
Pieschl, S. (2009). Metacognitive calibration—an extended conceptualization and potential applications. *Metacognition and Learning*, *4*(1), 3-31.
Puncochar, J. M., & Fox, P. W. (2004). Confidence in Individual and Group Decision Making: When" Two Heads" Are Worse Than One. *Journal of Educational Psychology*, *96*(3), 582.
Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*, 64-70.