Learning and Instruction 100 (2025) 102215

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

Testing individual and group markers of collaboration in a team-based learning classroom

Y. H. Victoria Chua [a,*], Justin Dauwels [b], Preman Rajalingam [c], Chew Lee Teo [d], Suzy J. Styles [a,e]

[a] School of Social Sciences, Nanyang Technological University, 48 Nanyang Ave, 639818, Singapore
[b] Department of Microelectronics, TU Delft, Mekelweg, 4 2628 CD, Delft, the Netherlands
[c] Institute of Learning, Mohammed Bin Rashid University of Medicine and Health Sciences, Building 14, Dubai Healthcare City, P.O Box 505055, Dubai, United Arab Emirates
[d] Center for Research in Pedagogy and Practice (CRPP), National Institute of Education, 1 Nanyang Walk, 637616, Singapore
[e] Centre for Research and Development in Learning (CRADLE@NTU), Nanyang Technological University, 61 Nanyang Drive, 637335, Singapore

ARTICLE INFO

Keywords:
Team-based learning
Collaborative discourse
Multi-party turn taking
Collaborative learning analytics
Co-located collaborative learning

ABSTRACT

Background: Intra-group discussions during actual TBL sessions play a huge role in knowledge consolidation and learning but are often understudied.
Aims: Using a pre-registered study framework, we examined if participation equity (H1), reciprocal interaction (H2), information density (H3), mutual understanding (H4), and emotional rapport (H5) affected how much students learn from their intra-group team-based learning discussions and how they rated their team's discussions.
Sample: Participants were 165 undergraduate students assigned to 28 teams.
Methods: Using linguistic, conversational, and socio-affective features extracted from recordings of Year 1 and 2 medical students engaging in team-based learning, each construct was conceptualised at the level of the group and the individual. We used linear mixed-effects models and competing models approach to establish which of our metrics best account for the observed variance in individual learning gains and perceived collaboration quality. The analysis plan was preregistered, including correction for multiple comparisons.
Results: None of our individual-level or group-level metrics significantly predicted individual learning gains. One of the group-level metrics significantly predicted perceived collaboration quality: reciprocal interaction. Our exploratory analysis found that individual baseline score of the best performer in the team positively predicted individual learning gains for others in their team, regardless of other interaction metrics.
Conclusion: While students perceived the highest collaboration quality when turn-taking in their team was evenly distributed, the strongest predicter of learning gains for a student was the knowledge level of their top-scoring team-mate. This finding has implications for classroom equity, group formation and activity planning.

1. Introduction

Collaboration is common and integrated into educational and work practices at all levels. As the use of collaborative practices in work and education have increased rapidly, it has become imperative to understand co-located collaborative experiences as well as its underlying processes. Despite extensive research on dyads, triads, and remote collaboration, realistic collaborative learning typically involves larger teams (Rajalingam et al., 2018), with some educators advocating for group sizes of four to six (Michaelsen et al., 2007; Parmelee et al., 2012). Furthermore, in many learning analytics studies, learners are observed outside of authentic classroom settings (Yan et al., 2023), and in such environments, collaborative behaviours of learners may well differ from the published literature.

One instance of large-group collaborative learning is team-based learning, which emphasises individual accountability and student

This article is part of a special issue entitled: Implementing Multimodal Learning Analytics published in Learning and Instruction.

* Corresponding author.
E-mail addresses: chua1005@e.ntu.edu.sg (Y.H.V. Chua), JHGDauwels@tudelft.nl (J. Dauwels), preman.rajalingam@dubaihealth.ae (P. Rajalingam), chewlee.teo@nie.edu.sg (C.L. Teo), suzy.styles@ntu.edu.sg (S.J. Styles).

https://doi.org/10.1016/j.learninstruc.2025.102215
Received 8 December 2024; Received in revised form 20 June 2025; Accepted 21 August 2025
Available online 18 September 2025

knowledge application through intra-group and inter-group discussions of problems designed to foster complex thinking and reasoning (Parmelee et al., 2012). The team-based learning process comprises three stages and begins with an individual preparation stage where students engage with learning resources prior to class. The subsequent stages occur in class where students are assigned to teams of five to seven (Parmelee et al., 2012; Rajalingam et al., 2018). During the second stage, the class session begins with students individually answering a set of closed-book multiple-choice questions (MCQs). These MCQs are designed to test whether the students have completed the pre-lesson study materials and give an overview on the foundational concepts of each lesson (Rajalingam et al., 2018). Students then repeat the same closed book MCQs in their assigned teams. During the intra-group discussion, students exchange knowledge and clarify questions and misconceptions (Parmelee et al., 2012; Rajalingam et al., 2018). The team submits one answer together and are permitted multiple attempts per question until correct. In the final stage, teams work on open-ended problem cases and participate in expert-facilitated class-wide inter-group discussions on the problem cases.

Observations of intra-team discussions during the second stage revealed variability in teams' capacity to provide clear and meaningful justifications, as well as arrive at accurate answers (Koh et al., 2020; Leupen et al., 2020). Yet, this critical aspect of team-based learning remains understudied. To understand the impact of authentic team dynamics during intragroup discussions on learning and teamwork perceptions, we conducted an examination of intra-group discussions among team-based learning teams of five to six students. Notably, this research was conducted in an ecological setting, where students were recruited from a medical school which uses team-based learning as its primary pedagogical approach during the first two years of the undergraduate medical program. Teams were recorded in real-time during actual classes, capturing their natural interactions and collaborative dynamics.

Following a pre-registered study protocol (Chua & Styles, 2023; https://osf.io/jb5a6), we examined if participation equity, reciprocal interaction, mutual understanding, information density and emotional rapport affected how much students learn from their intra-group discussions and how they rated their team's discussions. We capitalised on the multimodal nature of speech (Di Mitri et al., 2018; Chua, Dauwels, & Tan, 2019; Praharaj et al., 2021) by extracting a rich set of linguistic, conversational, and socio-affective features to examine our constructs of interest. Each construct was conceptualised at the level of the group and the individual, allowing us to investigate behaviours from a "person-plus-surround" perspective (Perkins, 1993). Using this strategy, we sought to understand how individual students benefit from different combinations of their own characteristics and the characteristics of their team-mates during authentic collaboration. In the following sections, we provide the theoretical framework for each construct, accompanied by hypotheses of each construct.

1.1. Participation equity

Given the zero-sum game of conversation, achieving participation equity (Cohen, 1994) – where all members have equal opportunities to contribute to and build upon the shared epistemic space – is crucial. Perceived imbalance in individual efforts negatively impact team cohesiveness, satisfaction and performance (Salomon & Globerson, 1989; Jassawalla et al., 2008; Teng and Luo, 2015). For example, triads with uneven speaking time distribution, where one member dominates the conversation, tend to be less effective than those with more even participation (Kapur, 2008; Borge & Carroll, 2014; Woolley et al., 2010).

In several learning analytics studies, participation equity is often defined as the variation in outputs among team members, with higher values reflecting less equitable distribution. Variances in individual speaking time (Bachour et al., 2010; Bergstrom & Karahalios, 2007; Borge & Carroll, 2014; Dowell et al., 2019; MacNeil et al., 2019; Shah et al., 2014; Woolley et al., 2010), utterance lengths (Lai et al., 2013), number of words produced (Vogel et al., 2023), speech rates (Koutsombogera et al., 2018), number of ideas produced by each team member (Tausch et al., 2014) or time spent interacting with task-relevant objects or shared devices (DiMicco et al., 2004; Marshall et al., 2008; Cukurova et al., 2020) have been operationalised as measures of balanced participation.

Despite the growing interest in encouraging and detecting participation equity, findings indicate that inequities in participation naturally emerge in ecological settings as individuals position themselves based on perceived competencies, expertise, or roles (Shah et al., 2014). Moreover, as team size increases, competition for participation and coordination costs increase (Holler et al., 2021). Consequently, optimal strategies for participation in larger groups may differ. For instance, it can be effective to designate experts to lead the discussion and achieve superior task outcomes (Cohen & Lotan, 2014; Lewis & Shah, 2015). Conversely, a team member may contribute sparingly but enhance the collaboration greatly by interjecting at right times (Strijbos & De Laat, 2010).

Given the critical role of participation in collaboration, we propose the following hypotheses (H1a and H1b) to explore the impact of participation equity on multi-party collaborative outcomes.

H1. Participation equity (as a relative proportion of the group's talk time) predicts both individual learning gains and perceived collaboration quality, in the following ways:

H1a. The more balanced participation is, the greater the individual learning gains will be. Individuals who participate more equally (or are working in teams with others who do so) will receive greater learning gains.

H1b. The more balanced participation is, the higher the perceived collaboration quality of team discussion will be. Individuals who participate more equally (or are working in teams with others who do so) will rate the collaboration quality of the team discussion higher.

1.2. Reciprocal interaction

Vygotsky's theory (1978) underpins collaborative learning, emphasising mutual engagement and knowledge sharing. Teams that engage in reciprocal interactions, where members frequently respond to one another, tend to achieve better collaborative outcomes. Frequent responding in online collaborative learning is linked to greater individual learning gains (Dowell et al., 2019), higher-quality collaborative products (Woolley et al., 2010; Zhou et al., 2020), and improved collaboration quality (Stewart et al., 2021). Within co-located collaborative learning scenarios, reciprocal interaction is often operationalised as speech overlaps (Bassiou et al., 2016; Zhou et al., 2014; Zhao et al., 2022; Kim et al., 2015), and interrupts (Oviatt et al., 2015) between students, and have been found to positively correlate with collaboration task success.

Turn-taking patterns in collaborative discourse can indicate information flows and knowledge asymmetries amongst team members (Damon & Phelps, 1989; Gibson et al., 2019; Popov et al., 2017). In comparison to dyadic interactions, multi-party turn-taking patterns can be more complex due to the growing number of possible turn transitions and knowledge asymmetries among team members, leading to varying degrees of reciprocal interaction. Multi-party intra-group discussions may alternate between periods of different students dominating the conversational floor and intense discussion involving multiple speakers (Haller et al., 2000; Gibson et al., 2019). Consequently, recent studies have adopted Shannon's informational entropy (Shannon, 1948) to measure variability in turn transitions during multi-party conversations (e.g., Lai et al., 2013; Wiltshire et al., 2021; Trujillo, 2024). Teams that have more distributed patterns of turn-taking across team members tend to be more satisfied with their teams (Koutsombogera & Vogel, 2019; Lai et al., 2013) and perform better (Haan et al., 2021).

Given the reciprocal nature of turn-taking, it is also important to understand where a student may benefit not from just engaging intensely with their team members but also by virtue of being in a team with intense reciprocal interactions. The critical role of reciprocal interaction in fostering effective collaboration underpins our hypotheses H2a and H2b, stated as follows.

H2. Reciprocal interaction in the team discussion will predict both individual learning gains and perceived collaboration quality, in the following ways:

H2a. The more reciprocal interactions occur during team discussion, the greater the learning gains observed.

H2b. The more reciprocal interactions occur during team discussion, the higher an individual will rate the collaboration quality of the team discussion.

1.3. Information density

When individuals collaborate, they balance the act of providing relevant yet sufficiently novel information during the team discussion. Collaborative discourse studies highlight that successful teams consist of team members who enhance previous contributions by providing new and diverse information (Kauffeld et al., 2018; Andrews-Todd & Forsyth, 2020) and focus more of their discussions on unique and complementary information among team members (Lu et al., 2012; Stasser & Titus, 2003). Given that it can be hard to manipulate the amount of information available to each student in authentic learning settings, the construction of collaborative learning groups tends to prioritise putting together members that optimise informational density through diversity in prior knowledge and individual preparation (Lou et al., 1996; Michaelsen et al., 2007; Parmelee et al., 2012; Zhang et al., 2016).

While information density may have positive effects on learning gains and team performance, perceived satisfaction of working in groups with differing levels of knowledge tend to be mixed (Haller et al., 2000). Studies on information density measured through lexical diversity reveal conflicting findings. While Reilly & Schneider (2019) found that dyads who exchanged contributions that are more lexically diverse had greater learning gains but lower collaboration quality ratings, Sinclar and Schneider (2021) found a positive relationship between lexical diversity and collaboration quality.

In all, these findings suggest that information density may have opposite effects on learning gains versus perceived collaboration quality. Thus, we formulate the following hypotheses H3a and H3b.

H3. Information density of verbal contributions exchanged in a team discussion will predict individual learning gains and perceived interaction quality, in the following ways:

H3a. The higher the information density of verbal contributions in the team discussion, the more learning gains the student will receive.

H3b. The higher the information density of verbal contributions in the team discussion, the lower the student will rate the collaboration quality of the team discussion.

1.4. Mutual understanding

For successful collaboration, sharing and maintaining a common ground plays a very important role (Roschelle & Teasley, 1995; Dillenbourg, 1999; Barron, 2003). Common ground is often conceptualised as the mutually shared knowledge, values, beliefs, and assumptions in a conversation within interlocutors (Clark & Brennan, 1991). Accounts of joint meaning-making suggest that mutual understanding is established through behavioural alignment (Pickering & Garrod, 2004; Rasenberg et al., 2020). Behavioural alignment refers to a large umbrella of phenomena related to behaviour matching (Rasenberg et al., 2020; Burgoon et al., 2017) and can be observable in linguistic acts (Pickering & Garrod, 2004), bodily gestures (Louwerse et al., 2012), and eye gaze (Wohltjen & Wheatley, 2021).

Given the diverse foci in existing literature, we narrow our focus to a specific linguistic mechanism of alignment, i.e., lexical entrainment, also known as lexical alignment. Lexical alignment is the process in which interlocutors re-use each other's words when speaking and has been argued to facilitate synchrony in interlocutors' mental models (Dale et al., 2013; Pickering & Garrod, 2004). Studies examining effects of lexical alignment on collaborative task performance reveal contrasting effects. Word re-use negatively predicted collaborative task performance (Heuer et al., 2020; Xu & Reitter, 2017) and observers' perceptions of team competence (Reverdy et al., 2020), while re-use of task-relevant words positively predicted task performance (Fusaroli et al., 2014; Fusaroli & Tylén, 2016). Studies examining collaboration quality reveal similar contrasting effects. Re-use of function words promotes trust (Gonzales et al., 2010) and perceived social support within the team (Heuer et al., 2020), but only among teams lacking a history of collaboration or high trust (Carmody et al., 2017). Altogether, these studies suggest that lexical alignment bears affective and cognitive functions in collaboration, where its cognitive function is to facilitate mutual understanding for goal achievement and its affective function is to facilitate social cohesion (Giles & Gasiorek, 2013; Burgoon et al., 2017).

The work on lexical alignment, at large, has been performed on dyads (see Pickering & Garrod, 2021 for a review), and studies on the effects and mechanisms of lexical alignment in larger groups are limited. A recent study (Duran et al., 2024) found that the degree of turn-by-turn lexical alignment within triads working on a physics task did not predict solution quality. Given the scarcity of research on lexical alignment beyond dyadic interactions, we investigate the impact of lexical alignment on multi-party collaborative outcomes with the following hypotheses.

H4. Establishing mutual understanding with team members through lexical alignment during the team discussion will predict both individual learning gains and perceived interaction quality, in the following ways:

H4a. The more lexical alignment is observed during team discussion, the more learning gains the individual will receive.

H4b. The more lexical alignment is observed during team discussion, the higher the individual will rate the collaboration quality of the team discussion.

1.5. Emotional rapport

Several frameworks used for analysing collaborative discourse emphasise the importance of establishing a positive socio-emotional climate and emotional rapport during collaboration (Kauffeld et al., 2018; Bakhtiar et al., 2018). Individuals and groups may enact various strategies, such as using humour, acknowledging one another's contributions, giving positive verbal encouragement, to regulate emotions, establish emotional rapport and maintain a positive socio-emotional climate (Linnenbrink-Garcia et al., 2011; Bakhtiar et al., 2018; Kauffeld et al., 2018). Collaborative learning studies have generally found the role of humour to be positive, improving engagement (Kim & Ho, 2018), fostering a sense of rapport and shared identity (Berge & Danielsson, 2013), and alleviating negative emotions associated with failure and frustration (Linnenbrink-Garcia et al., 2011; Lamminpää & Vesterinen, 2018).

In the current investigation, we focus on the role of joint laughter as a strategy for establishing emotional rapport during collaborative learning for the following reasons. First, unsurprisingly, laughter often indicates the presence of humour or enjoyment (Bachorowski & Owren, 2001; Platt et al., 2013). Second, shared laughter has been conceptualised as an intentional product of coordination and joint attention during conversations (Kangasharju & Nikko, 2009; Holt, 2010). The uptake of shared laughter between interlocutors can be similar to the ceding and receiving of speech turns, indicative of a form of conversational coordination, resulting in shared positive affect and possibly fostering supportive team communication patterns (Goodwin & Goodwin, 2012; Strid & Cekaite, 2021). In all, shared laughter not only indicates a shared positive emotion, but may also indicate joint attention.

While studies have established the positive effects of reciprocal laughter on group affiliation and cohesion (e.g., Berge & Danielsson, 2013; Berge, 2017; Ponton et al., 2018), there is an observable lack of consensus on the effects of reciprocal laughter on task performance, with studies showing both positive (Gao et al., 2023) and negative (Wang et al., 2016) effects. To investigate the effects of emotional rapport through shared laugher more closely in an authentic multi-party collaboration context, we had the following hypotheses.

H5. Emotional rapport may predict individual learning gains and will predict perceived collaboration quality, in the following ways:

H5a. As laughter may have mixed outcomes on task performance, we predict that emotional rapport will affect learning gains, but the direction of the effect may be positive or negative.

H5b. The more emotional rapport is observed during team discussions, individuals will perceive higher collaboration quality from the team discussion.

1.6. Research objectives

For the purpose of establishing whether any of the constructs predicted our target outcomes, we considered each construct as separate groups of hypotheses (H1 to H5). Fig. 1 outlines the overall methodology of the study, including control and outcome variables. The predicted relationships and the analysis plan were pre-registered prior to data linkage (Chua & Styles, 2023; https://osf.io/jb5a6). The conceptual structure for testing each construct follows the same logic. Each construct was investigated using a measure that describes individual behaviour, i.e., the individual-level metric, and an alternative measure that relates the team members' behaviours on that construct while discounting the individual student, i.e., the group-level metric. Table 1 summarises the five constructs, their metrics, and our hypotheses.

2. Method

2.1. Learning context

This project builds upon a dataset of audio-video recordings of students engaging in team-based learning at the Lee Kong Chian School of Medicine in Nanyang Technological University. Before the team-based learning session, the students individually studied assigned materials outside of class. The class session began with students individually completing a closed-book multiple-choice quiz delivered via a learning management system. Students did not receive feedback on their responses. Following the individual task, the team discussion portion of the class session commenced. During this time, each team answered the same quiz with a new feedback system. Each response triggered feedback which allowed revision for incorrect answers until only a binary choice remained. The intra-group discussion for the quiz was recorded.

Each TBL class session featured a distinct quiz focused on medical topics, including histology, immunology, and gastrointestinal system. The number of questions per quiz varied across sessions (Median = 26, Range = 19–48). Each quiz was designed to assess foundational knowledge relevant to the session's topic and was calibrated to be moderately challenging, encouraging students to engage in peer discussion, clarify misconceptions, and deepen their understanding of core concepts (Parmelee et al., 2012; Rajalingam et al., 2018).

At the end of the team discussion, students provided demographic information and used Teamwork Quality Survey (Weimar et al., 2017) to rate the team discussion they just completed. The study was reviewed and approved by the host university (IRB-2017-07-038). All recordings were conducted by the first author with the help of two research assistants.

2.2. Recording set up

Fig. 2 shows a simulated example of the audio and video camera set up. A 360-degree panoramic camera was set up in the middle of the table to collect individual facial recordings, and a lapel microphone was attached to each student to collect audio recordings. Additional details on the set up and recording parameters are outlined in the online Supplemental S1.

2.3. Participants

The participants were Year 1 and 2 medical students from Nanyang Technological University. They were recruited through email blasts sent to the student cohort by the course facilitator. All teams who signed up were recorded, with no exclusions. Each participant was reimbursed $20 for their time.

A total of 32 sessions were recorded from teams of five to six students (N = 189 students). Two of these sessions (N = 12 students) were pilot sessions used for the development of transcription protocols and study metrics and were excluded for analysis (Chua & Styles, 2023; Chua et al., 2019). Pilot sessions were conducted in a quiet room, with only one team recorded at a time. The room was designed to closely resemble the learning studio in which the actual TBL sessions were held (Fig. 2). The remaining 30 sessions were obtained during regular TBL classes conducted between September 3, 2019 to October 11, 2019 across 12 class sessions. These sessions took place in the learning studio, which accommodated up to 25 teams, each consisting of five to six students. Depending on participant availability, between one and three teams were recorded per class session.

Of the 30 sessions, one session was from a team (N = 6 students) who re-participated in the study and was excluded from analysis. One session

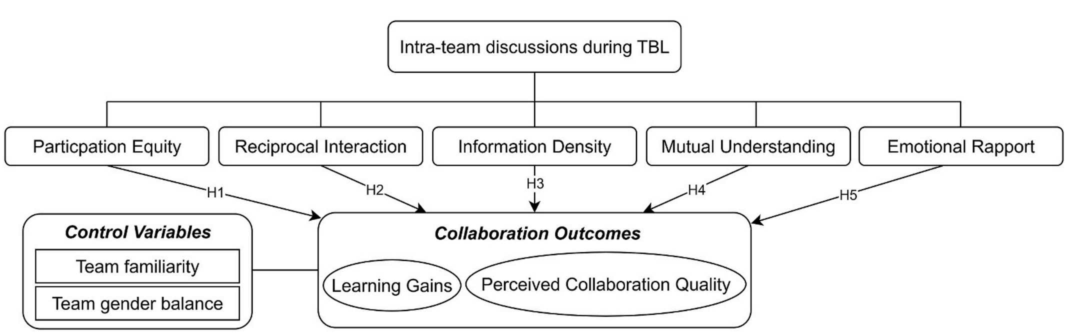

Fig. 1. Overview of methodology for investigating multiple constructs in this study.

Table 1
Table of Preregistered Hypotheses (H1 to H5), Constructs and Measures.

Construct	Measure	Individual measure	Group measure	Direction of Relationship
H1: Participation equity	H1a Learning gains	Individual absolute balanced participation index	Group absolute balanced participation index	+
	H1b Perceived collaboration quality	Individual absolute balanced participation index	Group absolute balanced participation index	+
H2: Reciprocal Interaction	H2a Learning gains	Individual turn-taking entropy	Group turn-taking entropy	+
	H2b Perceived collaboration quality	Individual turn-taking entropy	Group turn-taking entropy	+
H3: Information Density	H3a Learning gains	Individual lexical diversity	Group lexical diversity	+
	H3b Perceived collaboration quality	Individual lexical diversity	Group lexical diversity	–
H4: Mutual Understanding	H4a Learning gains	Individual lexical alignment level	Group lexical alignment level	+
	H4b Perceived collaboration quality	Individual lexical alignment level	Group lexical alignment level	+
H5: Emotional Rapport	H5a Learning gains	Individual co-laughing rate	Group co-laughing rate	+ or -
	H5b Perceived collaboration quality	Individual co-laughing rate	Group co-laughing rate	+

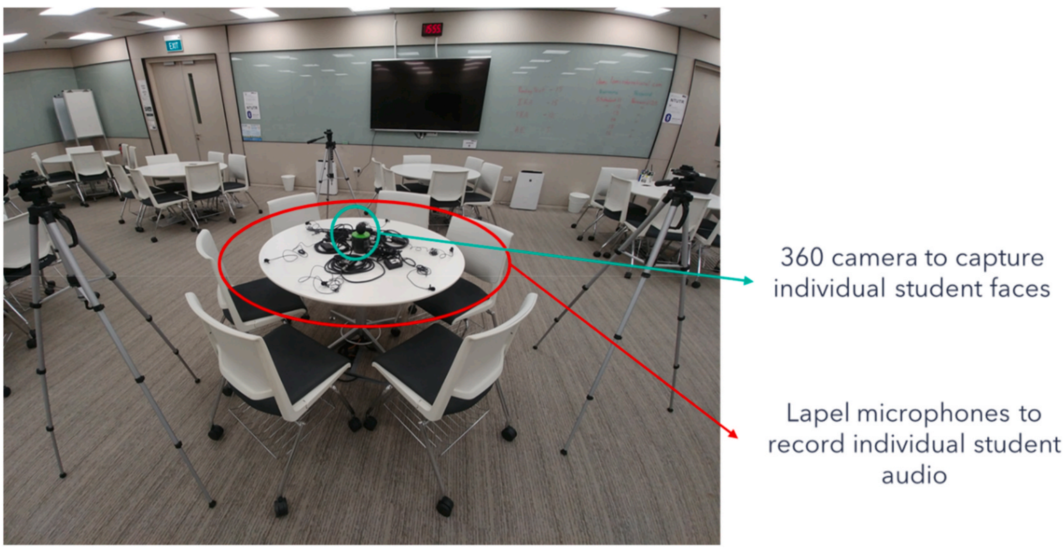

Fig. 2. Simulation of recording set-up.
Note: This photo was taken in a pilot study where one group was recorded in a controlled environment. Cameras on tripods were not included where recordings were performed in-class.

(N = 6 students) was excluded due to equipment failure resulting in audio data loss. In total, 28 sessions from 28 unique teams of 165 students (Female: N = 58; Male: N = 107), were included for analysis.

2.4. Audio data pre-processing and transcription

The intra-group discussions varied in length as they were recorded over different class sessions (Median = 11.2 min, Range = 6–42.1 min). The entire audio dataset was 17.82 GB. All intra-group discussions for the closed-book quiz were recorded and transcribed offline using a time-aligned transcription protocol. Current automatic speech recognition (ASR) systems struggle with various challenges, including multi-party conversations in noisy environments with overlapping speech (Watanabe et al., 2020), Southeast Asian accents (Chan et al., 2022), language mixing (Styles et al., 2023) and accurately capturing short words or discourse particles (Russell et al., 2024; Styles et al., 2023). Given that our dataset contains these features, we chose manual transcription, prioritising the reliability and validity of our findings, with the understanding that our findings would remain relevant in the context of future ASR advancements. A total of 9781 utterances were transcribed from 165 students in the dataset.

We only relied on the modality of audio to extract behaviours relevant to the five constructs outlined. As other studies have outlined (e.g., Spikol et al., 2017a, 2017b; Subburaj et al., 2020; Cukurova et al., 2020), there are potential interaction traces that can be extracted from the video modality. However, in our case, upon inspection of the video recordings, the faces of the students were sometimes blocked by their laptops and tablets. This is because the lesson plan was managed and administered via a learning management system and these devices were required for students to complete the lesson. As such, video data was not analysed in this paper and only used for providing additional context for transcription of audio recordings.

2.5. Measures

2.5.1. Outcome variables

Proportional individual learning gains. The individual and team answers to the multiple-choice quiz were captured by the learning management system that administered the questions. In our dataset, student teams were recorded in different weeks where instructors set different number of questions each week. Thus, to measure individual learning gains fairly across weeks, we used proportional learning gains (Hake, 2002), which we computed from the maximum possible score, the individual baseline score, and the team score, as outlined in

Equation 1.1

$$\frac{\textit{team score} - \textit{individual baseline score}}{\textit{maximum possible score} - \textit{individual baseline score}} \quad (1.1)$$

The maximum points possible for each question is computed as the number of options for each question minus one. For a question with five options, a maximum of four points can be awarded. The maximum possible score for the quiz is computed as the total number of questions in the quiz multiplied by maximum points possible for each question. As students can only attempt each question once during the individual portion of the class, the individual baseline score is computed as the number of correct answers given individually multiplied by maximum points possible for each question. During the intra-group discussion, teams are allowed to attempt the questions multiple times, and a point is deducted for each wrong attempt until no points are awarded if the team has attempted all options except the correct one. Thus, the team score is computed as the sum of points awarded for each correct group answer by the end of team discussion. The denominator indicates the maximum possible improvement the student could gain from the team discussion.

For students with a maximum individual baseline score, the denominator is the maximum possible score to account for potential loss where the team score is lower than the individual baseline score, as seen in Equation 1.2

$$\frac{\textit{team score} - \textit{individual baseline score}}{\textit{maximum possible score}} \quad (1.2)$$

Detailed worked examples are provided in the Supplement S2 to illustrate the computation of proportional individual learning gains.

Perceived Collaboration Quality. Perceived collaboration quality was measured using the Teamwork Quality Survey (Weimar et al., 2017) administered at the end of the team discussion. We did not use Evaluation of Team Performance as the students have the actual scores and feedback of their team's performance during the team discussion. Following Weimar et al. (2017), we summed the scores of Communication, Mutual Support, Cohesion, Trust, Shared Values, Coordination of Expertise to create the Perceived Collaboration Quality score for each student.

2.5.2. Control variables

Team size and the week of semester was documented at the time of recording.

Gender Balance in Team. Team gender composition was determined using the self-reported gender information provided by the students. A team was considered Male-Dominant if it had four or more male students, and Female-Dominant if it had four or more female students. Teams with an equal or near-equal gender distribution (e.g., 3:3 in teams of 6 or 3:2 in teams of 5) were classified as Balanced.

Team familiarity. Team familiarity was defined as the number of weeks the team has worked together since the start of the semester (T0). Teams recorded during the first week of the semester were coded as 0, the following week as 1, and so on.

2.5.3. Predictor variables

The individual-level and group-level metrics of all five constructs were derived from the transcripts.

Participation Equity (H1). Participation equity was measured using each student's speaking duration during the team discussion, defined from the start of quiz attempts to the submission of the final attempt, excluding off task talk. Individual speaking duration included overlaps with other team members, and total team discussion duration was the sum of all individuals' speaking durations, excluding pauses between speakers. Proportion of speaking time per student was individual speaking duration divided by total team discussion duration. Table 2 shows a worked example of the individual and group-level metrics computed for a group of five students. Interested readers are directed to Supplement S3 for more worked examples of these metrics.

Table 2
Example of individual and group metrics of participation equity (H1) in a group of 5 students.

Team discussion time (min)	Equal-share duration (min)	Student	Actual student talk time (min)	$\lvert\widehat{P}_A\rvert$	Total team discussion time minus student (min)	$\widehat{P}_{group\text{-}A}$
50	10	A1	0	1	50	1.6
	10	B1	5	0.5	45	2.22
	10	C1	10	0	40	3
	10	D1	20	1	30	2.67
	10	E1	15	0.5	35	2.86

$\lvert\widehat{P}_A\rvert$: Individual absolute balanced participation index, $\widehat{P}_{group\text{-}A}$: Group absolute balanced participation index.

Individual-level metric (H1): Individual absolute balanced participation index **($\lvert\widehat{P}_A\rvert$).** We used the absolute term of the balanced participation index from Dowell and colleagues (2019) to allow for comparison between groups of different sizes in our dataset, as shown in Equation 1.3.

$$\lvert\widehat{P}_A\rvert = \lvert n\overline{P}_{-}(A) - 1\rvert \quad (1.3)$$

Where.

- $\overline{P}_A$ is the proportion of speaking time of the student.
- n is the number of members in the team.

The individual absolute balanced participation index quantifies how balanced the individual's speaking duration is compared to what would be considered equal in a group of their size, irrespective whether participation is above or below what is considered equal. Specifically, an individual that speaks one unit more than the balanced amount has the same individual absolute balanced participation index of 1 as someone who speaks one unit less than the balanced amount. The higher the individual absolute balanced participation index, the more imbalanced the participation from the individual.

Group-level metric (H1): Group absolute balanced participation index **($\widehat{P}_{group\text{-}A}$).** For each student, the group absolute balanced participation index is computed as the sum of individual absolute balanced participation indexes of all other students excluding the current student of interest, as shown in Equation 1.4.

$$\sum_{b=1}^{n-1}\lvert(n-1)\overline{P}_b - 1\rvert \quad (1.4)$$

Consider the case of Student A1 (Table 2). To yield their group absolute participation index: we first calculated the total duration of team discussion excluding Student A. For each of the remaining team members, we computed their proportion of speaking time, using the adjusted total discussion duration as the denominator. Individual absolute balanced participation indexes for each of the remaining students were computed. These individual indexes were then summed to form the group absolute participation index for Student A. Similar to the individual-level metric, the higher the group absolute balanced participation index, the more imbalanced the group participation is across all other team members excluding the current student of interest.

Reciprocal Interaction (H2). The individual-level and group-level metrics of reciprocal interaction were derived from the students' turn-taking behaviours during team discussion.

Turns. To quantify turn-taking behaviours in a multi-party interaction, we outlined rules for multi-party turn taking. Following Duncan's (1972) criteria, utterances must be involved in a changeover of speakers (i.e., floor transfers) to be considered a turn. Not all utterances are involved in a turnover of the conversational floor. Vocal behaviours such as backchanneling and false starts were not considered turns and were excluded for analysis in this study. These rules were implemented via custom Python scripts. A detailed description of our turn-taking model is enclosed in Supplement S4.

Entropy-based measures. For each student, we used two entropy-based measures, reply entropy and response entropy. The entropy measures assessed the variability in the students' turn transitions when receiving responses (response entropy) and giving replies (reply entropy). Response entropy ($H\ (X_b\ |A)$) is calculated using the conditional probabilities of a student receiving responses from team members each time they speak. Higher response entropy indicates a more balanced distribution of responses across all team members. Thus, maximum possible response entropy ($H_{max}(X_b \mid \mathrm{A})$) is observed when responses to a student are evenly distributed among all peers. Conversely, reply entropy ($H\ (A\ |X_b)$) is calculated using the conditional probabilities of a student replying to each team member. Higher reply entropy indicates a more balanced distribution of the student's replies to all team members. Thus, maximum possible reply entropy ($H_{max}(\mathrm{A} \mid X_b)$) is observed when the student replies equally to all other team members. The individual-level and group-level metrics of reciprocal interaction were computed based on the reply and response entropies of each student as described in Equations 2.1 and 2.2.

Individual-level metric (H2): Individual turn-taking entropy [H (X_A)]. For each student, individual turn-taking entropy $H(X_A)$ was calculated as the sum of their response entropy and reply entropy as shown in Equation 2.1.

$$H(X_A) = \frac{H\ (X_b\ |A)}{H_{max}(X_b \mid \mathrm{A})} + \frac{H\ (A\ |X_b)}{H_{max}(\mathrm{A} \mid X_b)} \tag{2.1}$$

Where:

• $H\ (X_b\ |A)$ is the observed entropy based on the actual proportions of responses the student (A) received from every other team member (X_b), i.e., observed response entropy;

• $H_{max}(X_b \mid \mathrm{A})$ is the maximum possible entropy if the student received responses from every other team member in equal proportions;

• $H\ (\mathrm{A} \mid X_b)$ is the observed entropy based on the actual proportions of replies the student initiated towards every other team member, i.e., observed reply entropy;

• $H_{max}(\mathrm{A} \mid X_b)$ is the maximum possible entropy if the student initiated replies to every other team member in equal proportions.

The higher the individual turn-taking entropy, the more distributed the student's interactions are among their team members, indicating a higher level of reciprocal interaction exhibited by the student. To illustrate, Table 3 shows the turn transitions and conditional probabilities of a student with the lowest individual turn-taking entropy and a student with the highest individual turn-taking entropy in the dataset.

Group-level metric (H2): Group turn-taking entropy [$H(X_{group\text{-}A})$]. For each student in the team, we compute a group turn taking entropy that excludes the contributions by the student. We first exclude all turns by that student in the team discussion. The remaining turns amongst all other team members are used for the computation of group turn-taking entropy for that individual. For each individual, group turn taking entropy is computed using Equation 2.2.

$$H(X_{group-A}) = \frac{\mathrm{H_{group-A}}}{H_{(group-A)max}} \tag{2.2}$$

Where.

- $\mathrm{H_{group\text{-}A}}$ is the observed group turn-taking entropy for all turns occurring between each pair of speakers in the group, excluding the current student;
- $H_{(group-A)max}$ is the maximum possible group turn-taking entropy if all turns between each pair of speakers in the group, excluding the current student, occurred in equal proportions.

The higher the group turn-taking entropy of a student, the more distributed the interactions are among all other team members, indicating that the student is in a team with high reciprocal interaction. Interested readers can refer to Supplement S5 for further technical details of the individual-level and group-level metrics of reciprocal interaction.

Information Density (H3). To measure information density in the team discussion for each student, we define an individual-level metric (i.e., individual lexical diversity) from the words used by the student and a group-level metric (i.e., group lexical diversity) from the words used by everyone else in the team except the student. Lexical diversity is measured using the hypergeometric distribution D (HD-D: McCarthy & Jarvis, 2010). The HD-D estimates a student's lexical diversity by calculating the average contribution of each word type to the type-token ratio in a random sample of 42 words from their speech. Using the hypergeometric distribution, HD-D accounts for the varying likelihood of sampling frequent versus rare words (e.g., "a" vs. "catastrophic"), with frequent words contributing more to the ratio and resulting in lower HD-D scores. A fixed sample size of 42 tokens, as recommended by McCarthy & Jarvis (2010), is used to enable cross-study comparability. HD-D is computed using the *lexical_diversity* Python package (Kyle, 2023).

Individual-level metric (H3): Individual lexical diversity. For each student, the individual lexical diversity is computed as the HD-D of all words spoken by that student during the team discussion. A high score on this metric would indicate that the student uses a large variety of different words during the team discussion.

Group-level metric (H3): Group lexical diversity. For each student, group lexical diversity of other team members is computed as the HD-D of all words spoken by the group excluding the words spoken by that student during the team discussion. A high score on this metric would indicate that the student is in a group where the other team members are using a large variety of different words.

Mutual Understanding (H4). Mutual understanding is measured via the level of lexical alignment between contiguous speech turns

Table 3
Turn transitions of students with low and high individual turn-taking entropy.

Student	Individual turn-taking entropy	Response from:	No. of responses received	Probability	Replies to	No. of replies given	Probability
A1	1.2	B1	4	0.4	B	3	0.33
		C1	3	0.3	C	4	0.44
		D1	3	0.3	D	2	0.22
		E1	0	0	E	0	0
		F1	0	0	F	0	0
A2	1.78	B2	18	0.250	B2	16	0.216
		C1	15	0.208	C1	14	0.189
		D1	11	0.153	D1	16	0.216
		E1	15	0.208	E1	11	0.149
		F1	13	0.181	F1	17	0.230

Note: Student A in group 1 has the lowest individual turn-taking entropy and Student A in Group 2 has the highest individual turn-taking entropy in the dataset.

during the team discussion. We used the ALIGN Python package for lexical alignment computation (Duran et al., 2019). Only pairs of successive turns identified by our turn-taking model (Supplement S4), i.e., utterance-response pairs, were computed for lexical alignment. For each utterance-response pair, all words in both turns were lemmatised (i.e., 'apples' becomes 'apple'). Each utterance-response pair was represented as a pair of word vectors that accounts for the number of occurrences of each lemma or token. The ALIGN package computes lexical alignment scores based on cosine similarity between lemmas and tokens in utterance-response pairs. The higher the number of same lemmas or tokens present in the utterance-response pair, the higher the cosine similarity of the utterance-response pair, and hence the higher the lexical alignment scores. The lemma lexical alignment score and the token lexical alignment score are averaged to create a single lexical alignment score for each utterance-response pair.

Based on recommendations by Dideriksen and colleagues (2023), only utterance-response pairs that showed alignment were used for computing individual and group-level metrics in this study. For a more detailed implementation of our lexical alignment computation, we point interested readers to Supplement S6.

Individual-level metric (H4): Individual Overall Lexical Alignment Level. For each student, we computed their individual lexical alignment to others. This is defined as the average of lexical alignment scores of all aligned utterance-response pairs containing the student responding to someone. We also computed the individual lexical alignment from others, defined as the average of lexical alignment scores of all aligned utterance-response pairs containing the student receiving a reply from someone. Thus, the individual overall lexical alignment level is computed as the average of the student's individual lexical alignment level to others and their lexical alignment level from others. A high score would indicate that the student is aligning to their team members to a great extent.

Group-level metric (H4): Group Lexical Alignment Level. For each student, the group lexical alignment level is computed as the average of lexical alignment scores of all aligned utterance-response pairs in the team discussion that does not include that student. A high score on this metric would indicate that the student is in a group where other team members are aligning to a great extent to one another, regardless of whether they are aligning to the student.

Emotional Rapport (H5). Emotional rapport is measured via co-occurrences of laughter during the team discussion, i.e., laughter that occurs between two or more speakers during the team discussion.

Laughter coding. In the transcripts, the start and end times of individual laughter segments were manually identified. Whenever possible, laughter was separated from the main utterance. For instances where laughter co-occurred with speech, the laughter boundary was marked as beginning when the speech ceased. Transcribers consulted video recordings to identify the speaker in cases where the source of laughter was ambiguous.

Co-laughing. Adopting the perspective that initiation of laughter from the first speaker and reciprocal laughter from another speaker is reminiscent of turn-taking (Dupont et al., 2016), we define shared laughter, i.e., co-laughing, with the following logic. For every instance of an individual's laugh, an instance of another individual laughing occurring within 1500ms or less at the end of the speaker's laugh is marked as co-laughing. A sequence of co-laughing is defined as starting when the first student begins laughing, continues when one or more students laugh at the same time, or in quick succession after another in the sequence (threshold = 1500ms), and concludes when the last student in the sequence stops laughing. Fig. 3 shows an example of a co-laughing sequence shared by three students. These rules were implemented via custom Python scripts. To measure emotional rapport in the team discussion, we define an individual-level metric (i.e., individual co-laughing rate) and a group-level metric (i.e., group co-laughing rate).

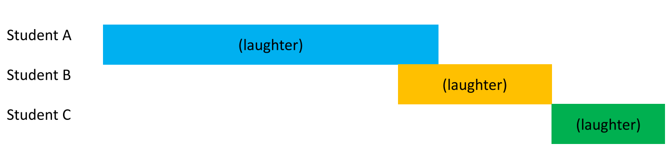

Fig. 3. *Visualisation of a Co-Laughing Sequence.*
Note: Each horizontal bar represents the duration of laughter and is coloured by the student. All laughs depicted were considered shared.

Individual-level metric (H5): individual co-laughing rate. The individual co-laughing rate is computed as the number of times a student laughs with someone else in their group during the team discussion divided by the total duration of the team discussion. A high score on this metric would indicate that the student does a lot of co-laughing with their teammates, as would be expected if the student has good rapport with their teammates.

Group-level metric (H5): group co-laughing rate. For each student, the group co-laughing rate is computed as the number of times their team members laugh with each other without the student divided by the total duration of the team discussion. A high score on this metric would indicate that the student is in a group where the other team members do a lot of co-laughing without them, as would be expected if the other team members have good rapport with one another, regardless of whether the student shares this rapport.

2.6. Analytical approach

As students are nested within their teams, we performed linear mixed model analysis using the lme4 package (Bates et al., 2015) with R version 4.2.2. As students within the same group would have more similar scores than students from other groups (Cress, 2008; Janssen et al., 2013), we allowed the intercept term to vary across groups (i.e., random intercept) to account for the nonindependence in our dataset. Team familiarity and team gender balance were included as fixed effects for control.

Each construct was tested separately (Fig. 1). Tables S7.1 to S7.5 in Supplement S7 outlines the mixed-effects models and our covariate-only models for each construct (H1 to H5). We first checked if there was a significant main effect of the individual-level metric on the outcome variable of individual learning gains. We then checked if there was a significant main effect of the group-level metric on the outcome variable of individual learning gains. If both metrics were significant predictors and their correlation was smaller than 0.3, we combined them to predict the outcomes. We repeated this entire process for the outcome variable of perceived collaboration quality. To determine if individual- or group-level metrics predicted the outcome variable above and beyond the variance attributed to team familiarity, team gender balance and group characteristics, we conducted likelihood ratio tests between predictor models with significant main effects and a covariate-only model containing only team familiarity and team gender balance. As each construct might have involved a maximum of six models to be tested (three each for learning gains and collaboration quality respectively), alpha was controlled for the use of multiple models within each construct ($\alpha = .0083$, corrected for six models). Our planned and pre-registered analyses (Chua & Styles, 2023) were well-powered at 0.95. Further details on our analytical approach are documented in Supplement S7.

As it may also be valuable for researchers reading our work to determine which effects investigated in this study are small enough to be practically irrelevant, we conducted equivalence testing. Further documentation of our equivalence testing approach is included in Supplement S8.

3. Results

The correlations between outcome variables and all predictor variables (H1 to H5) and model outputs of all tested linear mixed models are enclosed in Supplement S9.

3.1. H1: Participation Equity

3.1.1. H1a: Does participation equity predict individual learning gains?

Neither participation equity measured at the individual level (β = 0.00236, t = 0.039, p = .969) nor at the group level (β = 0.0299, t = 0.868, p = .389) significantly predicted individual learning gains.

3.1.2. H1b: Does participation equity predict perceived collaboration quality?

None of our individual-level (β = 1.51, t = 0.441, p = .660) or group-level (β = −2.48, t = −1.75, p = .0879) metrics significantly perceived collaboration quality.

3.2. H2: Reciprocal interaction

As both individual-level and group-level metrics of reciprocal interaction were highly correlated (rho = 0.71, p < .001; Table S9.1), we did not fit models with both predictors.

3.2.1. H2a: Does reciprocal interaction predict individual learning gains?

Students appeared to have lower learning gains when they had higher individual turn-taking entropies (β = −0.467, t = −2.21, p = .0292), but this effect was not statistically significant at our preregistered alpha value of 0.0083. There was insufficient evidence to support the hypothesis that individual learning gains differed as a function of group turn-taking entropy (β = −0.168, t = −0.248, p = .805).

3.2.2. H2b: Does reciprocal interaction predict perceived collaboration quality?

Students appeared to have higher perceived collaboration quality when they had higher individual turn-taking entropies (β = 19.5, t = 2.13, p = .0382). However, this effect was not statistically significant at the preregistered alpha value of 0.0083. There was significant main effect of group turn taking entropy (β = 72.8, t = 2.85, p = .00729) on perceived collaboration quality scores. As seen in Fig. 4, individuals in groups that had higher turn-taking entropy rated their team discussion as better.

The model with group-turn-taking entropy as an additional predictor yielded a significantly better fit than the covariate-only model, $\chi2(1)$ = 8.14, p = .00433, with lower AIC and BIC values as seen in Table 4. However, our fitted model explained only a total of 7.5 % of the variance observed (conditional R^2). The fixed effects of our model explained 5.7 % of the variance observed (marginal R^2). The covariates explained a total of 0.8 % of variance observed (marginal R^2 of covariate model). As such, group turn-taking entropy explained only 4.9 % of the variance observed in our data.

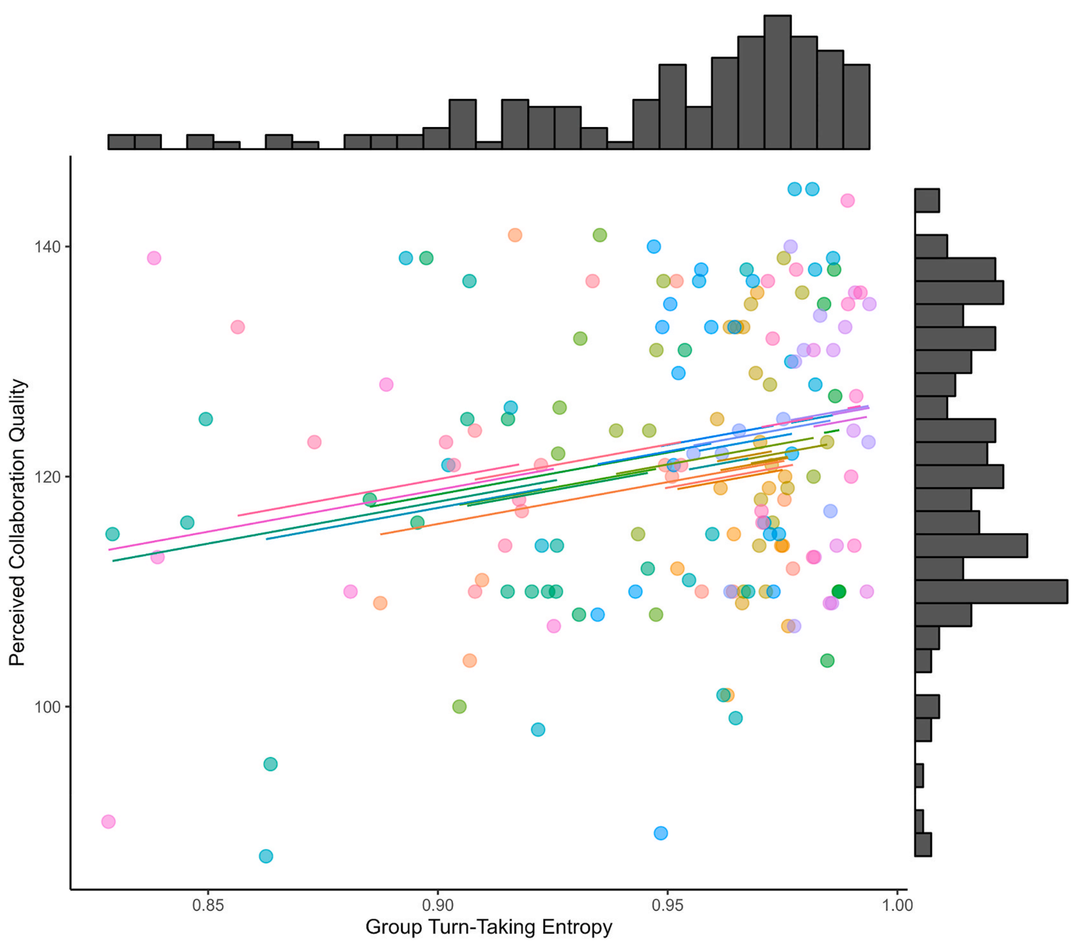

Fig. 4. Group turn-taking entropy and perceived collaboration quality scores.
Note: Scatterplot showing group turn-taking entropies (x-axis) and perceived collaboration quality scores (y-axis) for each student. The regression lines for predicting perceived collaboration quality scores with group turn-taking entropies is fitted for each group, allowing each group to have their own intercept. The covariates team familiarity and team gender balance are included in the model as well. Each dot represents one student and is coloured by the group the student is in. The regression slopes are also coloured by group. Alpha of dots and lines is set to 0.5.

Table 4
Covariate-only model and full model with group turn-taking entropy as an additional predictor for perceived collaboration quality scores.

	Covariate-only model	Full model
(Intercept)	117.836**	48.344
	(3.929)	(24.635)
Team Familiarity	0.620	0.704
	(0.646)	(0.572)
Team Gender Balance	−0.227	−0.401
	(2.403)	(2.120)
Group Turn-Taking Entropy		72.770**
		(25.500)
SD (Observations)	11.875	11.872
Num.Obs.	165	165
R2 Marg.	0.008	0.057
R2 Cond.	0.075	0.075
AIC	1295.2	1281.6
BIC	1310.8	1300.3
ICC	0.1	0.0
RMSE	11.49	11.64

Note: *p < .05, **p < .0083.

3.3. H3: Information density

3.3.1. H3a: Does information density predict individual learning gains?

There was insufficient evidence to support the hypotheses that individual learning gains differed as a function of information density at the individual level ($\beta = 0.130$, $t = 0.266$, $p = .791$) or at the group level ($\beta = 0.834$, $t = 0.552$, $p = .583$).

3.3.2. H3b: Does information density predict perceived collaboration quality?

Neither information density measured at the individual level ($\beta = 14.9$, $t = 0.556$, $p = .579$) or at the group level ($\beta = 0.294$, $t = 0.005$, $p = .996$) significantly predicted perceived collaboration quality.

3.4. H4: Mutual understanding

3.4.1. H4a: Does mutual understanding predict individual learning gains?

Neither lexical alignment at the individual level ($\beta = 0.00236$, $t = 0.039$, $p = .969$) or at the group level ($\beta = 0.0299$, $t = 0.868$, $p = .389$) significantly predicted individual learning gains in the current task.

3.4.2. H4b: Does mutual understanding predict perceived collaboration quality?

There was insufficient evidence to support the hypotheses that perceived collaboration quality scores differed as a function of overall alignment level at the individual level ($\beta = 12.5$, $t = -0.599$, $p = .550$ and at the group level ($\beta = 16.5$, $t = 0.601$, $p = .551$).

3.5. H5: Emotional rapport

3.5.1. H5a: Does emotional rapport predict individual learning gains?

There was insufficient evidence to support the hypotheses that individual learning gains differed as a function of co-laughing rate at the individual level ($\beta = 0.001$, $t = 0.013$, $p = .989$), team familiarity ($\beta = -0.0191$, $t = -1.04$, $p = .309$) or at the group level ($\beta = -0.0346$, $t = -0.420$, $p = .675$).

3.5.2. H5b: Does emotional rapport predict perceived collaboration quality?

Shared laughter measured at the individual level ($\beta = 0.270$, $t = 0.102$, $p = .919$) or at the group level ($\beta = -2.57$, $t = -0.660$, $p = .512$) did not significantly predicted perceived collaboration quality.

3.6. Equivalence testing

Our equivalence tests revealed that of all predictors of interest in this study, only the effect of participation equity (H1) measured at the group level on learning gains is too small to be considered meaningful in our study (Supplement S8).

3.7. Exploratory analysis

As shown in Fig. 5, we observed that teams with high scoring members made higher gains, while teams with lower scoring members made modest gains. To understand how students with different baseline knowledge benefit from collaboration, we conducted an exploratory investigation of all learners with the most to gain from collaboration.

3.7.1. Exploratory analysis methods

We performed the analysis on 128 students, excluding the highest scoring students in each group. For each of the constructs: participation equity, reciprocal interaction, mutual understanding, information density and emotional rapport, we followed the method described under Analytical Approach (for more details see Supplement S10). However, for all models, we added an additional fixed effect term of the individual baseline proportion score from the highest performer in each group.

3.7.2. Exploratory analysis results

In the interest of brevity, we only discuss significant effects when observed (all results enclosed in Supplement S10). We observed that the individual baseline score of the best performer in the team positively predicted individual learning gains for others in their group regardless of other constructs investigated (e.g., $\beta = 0.594$, $t = 2.53$, $p = .018$), but this effect was not significant at our pre-registered alpha value. As the analytical approach resulted in a smaller sample size (N = 128), the exploratory analysis was underpowered to detect our effect size of interest (Figure S10.1 in Supplement S10. We caution that the effect should be replicated in future studies before it is treated as a reliable effect. Perceived collaboration quality scores differed as a function of reciprocal interaction at the group level ($\beta = 94.5$, $t = 2.89$, $p = .00649$). The direction of this effect is positive, as previously observed.

4. Discussion

Previous studies on collaborative learning have established several constructs believed to impact teamwork and learning outcomes. However, few studies have investigated these dynamics in ecological large group teaching approaches. Thus, we investigated the impact of participation equity (H1), reciprocal interaction (H2), mutual understanding (H3), information density (H4), and emotional rapport (H5) on outcomes of learning gains and perceived teamwork quality within an ecologically valid context of team-based learning and focused on an understudied aspect of team-based learning: intra-team discussions. Although this paper focuses on the analysis of audio recordings, we harnessed multiple streams of data from speech and extracted linguistic, conversational, and socio-affective features to examine our constructs of interest. To our knowledge, our investigation is the first of its kind to examine collaboration dynamics in large groups by separately considering the individual's contributions and contributions from their other team members within the same team discussion. The current investigation can be best understood as an exploratory design that was pre-registered.

None of our individual-level or group-level metrics significantly predicted individual learning gains after correcting for multiple comparisons (H1-5a). When testing H2a, we observed that individual turn-taking entropy had a negative relationship with learning gains at the standard level of alpha, p < .05. In our equivalence tests, this contrast could not be accepted as equivalent (Supplement S8), indicating there is insufficient data to conclude whether the observed data is similar to or

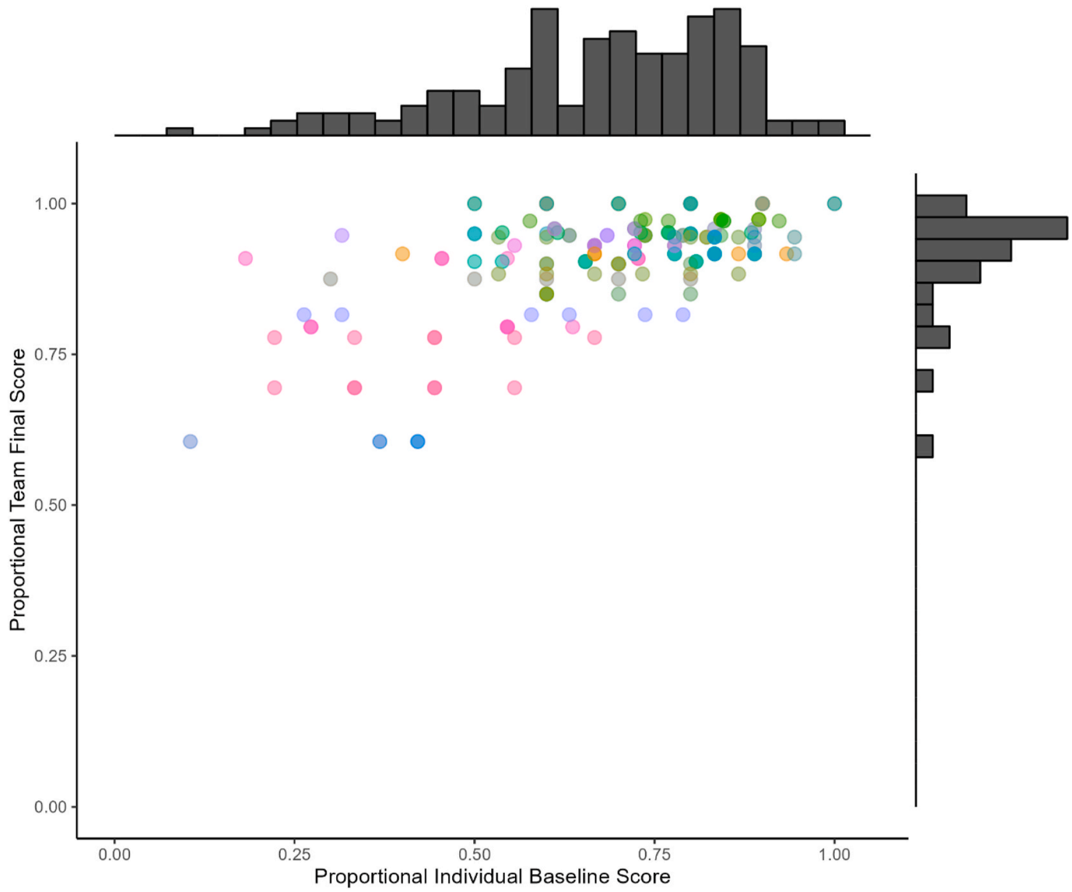

Fig. 5. Individual baseline scores and team scores.
Note: Scatterplots showing proportional individual baseline scores (y-axis) and proportional team scores for each student (x-axis). Each dot is coloured by the group the student is in. Proportional individual baseline scores are computed as individual baseline scores divided by maximum possible score of the task. Proportional team scores are computed as team scores divided by maximum possible score of the task. Alpha of dots is set to 0.5.

different from what would be observed by chance. It is possible the true effect size of individual turn-taking is smaller than the minimum effect size of interest, and our study is underpowered to detect a true effect. A post-hoc power analysis suggests that future investigations would require a minimum sample size of 341 to detect the observed effect size of 0.0384 we found in this study (Figure S9.1 in Supplement S9). A single year-level cohort in the current context of our study is around 150 students (Nanyang Technological University, 2024), and future researchers should keep in mind that potential replication studies may require recruitment from a larger participant pool (e.g., cross-institutional). Conversely, the measurement of individual turn-taking entropy may not be ideal for measuring reciprocal interaction, and alternative metrics may be more suitable to capture the construct. In all, this effect should be treated with caution until replicated reliably in future samples.

In terms of collaboration quality, only the group-level metric of reciprocal interaction, group turn-taking entropy, positively predicted perceived collaboration quality scores (H2b). The individual-level metric of reciprocal interaction was not a significant main effect of collaboration quality at our pre-registered alpha level of $p < .0083$. In our study, the experience of being in a team with high reciprocity led students to perceive their teams as more effective, irrespective of individual engagement and individual learning gains. This finding partially aligns with prior research, supporting evidence that balanced turn-taking enhances team cohesion and belonging (Hu & Chen, 2021; Koutsombogera & Vogel, 2019; Lai et al., 2013), but contrasts with studies linking it to improved learning outcomes (Dowell et al., 2019; Haan et al., 2021; Woolley et al., 2010). Educators implementing team-based learning in their classrooms should be aware that teams who appear to have the best working dynamic may value their interactions most but may not be the ones gaining the most from the task.

Finally, our exploratory analysis revealed that the highest performing individual affected the learning gains of the remaining team members, indicating that the role of high performing individuals should be considered more thoroughly when evaluating collaborative dynamics. This finding provides tentative support for existing studies illustrating that individuals benefit differentially in collaboration; specifically, lower performing students benefit from collaborating someone who is performing better than them (Deiglmayr & Schalk, 2015; Zambrano et al., 2019).

5. Limitations

Given the absence of significant differences in collaborative behaviours between individuals with higher and lower learning gains, we note important differences in the implementation of our studies that may have contributed to this disparity. First, our sample was limited to voluntary sign-ups, and there may have been self-selection biases in the sample. Students who did not enjoy collaborating with their group, and groups that did not enjoy working together may not have signed up for the study. This self-selection model may have reduced the diversity of teamwork dynamics, making it hard to investigate the full range of collaborative behaviours and their outcomes on learning.

Second, the measure of learning gains provides a limited view on the learning experienced by the student. The pedagogical model did not include post-collaboration tests of recall to assess individual knowledge retention. We were also unable to discount practice effects, as students worked on the same set of questions individually and then with their team. Some studies use the end-of-semester score as an outcome measure (Zilvinskis, 2019). However, the team-based learning sessions included in the current sample provided only a single snapshot of team-working behaviours from one week of class. It is unclear whether the learning gains from a single session would be applicable to the entire curricular content. In all, these limitations were inherent to the ecological context of our study, highlighting the trade-off between ecological validity and research rigour.

Third, the structured nature of our task, which had a single correct answer for each question in the quiz, may have limited collaborative behaviours shown to be beneficial in less constrained contexts. In such well-defined tasks, it can be more effective for teams to defer to the most knowledgeable member (Webb, 1995; Dillenbourg, 1999), which may account for our exploratory finding that teams with high-performing individuals achieved greater gains. In contrast, open-ended or information-distributed tasks benefit more from knowledge sharing and negotiation among team members (Aronson, 1978; Barron, 2003; Stasser & Titus, 2003). It is important to note that we are not implying that team-based learning paradigms like the one used in our study are ineffective for collaborative learning. Rather, the administered quizzes consisted of questions with varying levels of difficulty, which likely requires different trajectories of interaction at different times for successful completion, as noted in previous studies (e.g., Haller et al., 2000; Kutnick & Blatchford, 2014; Radebe & Mushayikwa, 2023). Hence, a potential future direction is to delineate portions of the team discussion where students are attempting a question multiple times and examine if the collaborative behaviours of interest affected the performance of different groups during completion of those questions.

Fourth, the multimodal proxies investigated in this study may not have been ideal for measuring our constructs. We relied on a single dimension to represent each construct. Take for example, emotional rapport. Like the other constructs in our study, emotional rapport is a complex, multidimensional construct involving various verbal and nonverbal strategies, including humour, verbal encouragement, and mutual acknowledgment (Linnenbrink-Garcia et al., 2011; Bakhtiar et al., 2018). In the current investigation, we only used co-laughing as its sole indicator. Although, shared laughter is recognised as an important social cue contributing to emotional rapport (H5) and a positive socio-emotional climate (Kauffeld et al., 2018; Bakhtiar et al., 2018), relying solely on it may not fully capture the role of laughter in ecological collaboration. Laughter may have functions beyond positive affect (Dupont et al., 2016), e.g., to indicate topic change (Koutsombogera & Vogel, 2022), end of a speech turn (Holt, 2010), or reduce speaker's embarrassment, discomfort, or anxiety (Jefferson, 1984; Wang et al., 2016). Studies interested in laughter as a proxy for emotional rapport should complement laughter annotations with qualitative coding to uncover laughter's nuanced functions in regulating emotional rapport within collaborative settings.

Finally, this study relied solely on audio data to extract collaborative behaviours. Although video can provide additional interaction cues, in our case, faces of the students were often obscured by laptops and tablets. Given the necessity of such devices and the increasing ubiquity of computer-mediated collaborative learning tasks in co-located educational settings, future studies may benefit from utilising built-in cameras on students' devices. Emerging studies suggest that high fidelity facial features can be successfully extracted from videos recorded by Zoom for understanding useful collaborative working behaviours (Zhou et al., 2020; Subburaj et al., 2020). Video conferencing software such as Teams and Zoom may also make it easier to sync and collect video recordings of students' faces in authentic learning settings.

In general, our results may reflect current limitations in using multimodal proxies to model learning outcomes. To provide a blueprint for future systematic inquiry into effective collaborative constructs, we have transparently reported all effects and uncorrected p-values (Supplement S9). Equivalence tests were also conducted (Supplement S8), indicating that of all our predictors of interest, only the effect of participation equity measured at the group level on learning gains should be treated as negligible. We hope this work informs the challenges in modelling collaborative learning from ecological settings and guides further investigation.

6. Implications

Contrary to our expectations, our findings suggest that strategies prioritising participation equity, while beneficial for dyads and small groups, may not necessarily translate to larger co-located groups. These findings resonate with prior research suggesting that natural inequities in participation can arise during collaborative activities due to task demands and student characteristics (Shah et al., 2014; Cohen & Lotan, 2014; Lewis & Shah, 2015). Future studies interested in participation should explore contexts where participation equity is desirable and those where it may not be necessary.

In the current investigation, we failed to observe significant effects of verbal activity on learning gains. Silence is usually treated as a feature for training models that predict negative collaborative behaviours (e.g., Cornide-Reyes et al., 2020; Vrzakov et al., 2020; Worsley et al., 2021), but the dynamics of silence may vary for students in a multi-party interaction especially when considering how silence may be distributed among members of a large group. According to previous studies, the decision to be silent can be a culturally appropriate act of respect for some groups of students (Remedios et al., 2012), a collaborative act of active listening (Poland & Pederson, 1998), or a strategic deference to a high-performing student in the team (Webb, 1995; Dillenbourg, 1999). Given the large number of learning analytics tools that emphasise verbal activity metrics, educators and learning analytic researchers should be cautious in judging low verbal activity as a sign of non-productive collaboration or disengagement. Educators considering implementation of such tools in their classrooms should be aware of potential biases and how they might lead to disparate impacts on student subgroups (e.g., based on cultural backgrounds and learning competencies) within their classrooms.

Given that audio data is a modality frequently collected in multimodal learning analytics studies (Chua et al., 2019), we recommend future research further explore other speech-based measures for turn-taking dynamics. While the directionality of turn-taking is important for understanding interpersonal coordination in collaboration, our findings suggest that entropy-based measures alone may be limited in capturing the complexity of these interactions. Alternative approaches, such as network analysis, may offer a more nuanced view of how turn-taking patterns relate to collaborative outcomes (Becheru et al., 2018; Claros et al., 2015). Additionally, investigating individual patterns, including the roles of silence and active listening, may reveal how participants regulate floor control and learning opportunities (Wise et al., 2012; Worsley et al., 2021). Moreover, burgeoning evidence suggest that it may be important to distinguish between turns addressed to the whole group or specific members (Gibson, 2003; Gibson et al., 2019) and how they are received in multi-party collaboration. For instance, students who tend to grab turns directed to another team member specifically have been found to be academically weaker (Hu & Chen, 2021). We posit that further investigation into such turn-taking patterns, accompanied with information on students' perceived competencies, expertise, and prior knowledge, can complement existing work on participation equity patterns and reciprocal interactions in collaborative learning contexts.

While students who have the highest individual baseline scores may not have been seen as experts in the group, in our task context, teams with more knowledgeable individuals achieved higher learning gains. We suggest that an equitable approach to group working is to ensure that team memberships change throughout the semester, so that all students have a chance to benefit from the highest performing students. In some circumstances, group composition could be integrated with pre-discussion test scores to ensure that peer-learning opportunities are optimised further based on prior knowledge for a specific task. Providing students opportunities to work with different groups throughout the semester prevents an entrenchment of students' perceived status in their group, potentially facilitating opportunities for different students to be a higher performing student in their assigned group.

Taken together, our current findings corroborate recent challenges faced by learning analytics tools when transitioning to novel settings (Haim et al., 2023; Kitto et al., 2023; Martinez-Maldonado et al., 2021). We suggest two recommendations to improve future implementation of ecological learning analytics studies. First, addressing the 'forking paths' problem: researchers face multiple choices during data collection, processing, and analysis, leading to unconstrained degrees of researcher freedom (Chambers, 2017). During our literature review, we noted a lack of guidelines in current literature for computing suitable multimodal proxies for each construct of interest. For instance, for mutual understanding (H4), we found diverse ways of computing lexical alignment in the literature. Different decisions may be made for type of words considered for analysis (e.g., all words, task-relevant words, function words, after removing high frequency words), scale at which lexical alignment is computed (e.g., turn-by-turn, window-based, whole conversation), or method for computing similarity (e.g., cosine similarity, Jaccard distance). Such variations increase the likelihood of unintentional methodological variations, highlighting how different researchers can develop distinct approaches in their investigations, leading to divergent conclusions (Gelman & Loken, 2013). Our findings support recent calls (Haim et al., 2023; Kitto et al., 2023) to standardise and document all data pre-processing and analysis decisions.

Secondly, we suggest researchers differentiate between exploratory and explanatory models when using behavioural features to predict collaborative outcomes (Gelman & Loken, 2013). Some researchers have suggested that learning analytic tools fail to replicate in other settings due to overfitting of models (Haim et al., 2023; Kitto et al., 2023). Problems in replicability may also be compounded by over-inflation of significance estimates uncorrected for multiple analyses. We demonstrate here that after correcting for multiple comparisons, we only found the significant effect of group turn-taking entropy on perceived collaboration quality (H2b). We agree with other researchers advocating pre-registration in this field (Ferguson & Clow, 2017; Haim et al., 2023; Kitto et al., 2023; Motz et al., 2022). Even pre-registering exploratory studies, as we have done here, can advance learning analytics, promoting the culture of research collaboration (Nosek et al., 2018), and avoiding the misinterpretation of exploratory results as confirmatory (Chambers, 2017; Kerr, 1998). Researchers should use exploratory findings to design follow-up confirmatory studies with pre-registered protocols. In this way, learning analytics research can achieve greater rigor and transparency, promoting the development of robust findings for improving educational practices.

7. Conclusion

As it becomes easier to collect data streams from co-located learning, it is important that we understand what types of behaviours are helpful in understanding co-located collaboration, and how these behaviours are related to meaningful constructs that facilitate collaboration. Our results emphasise the infancy of deploying learning analytics that leverage audio features in authentic learning settings. Reproducibility and generalizability are critical in the models developed for learning analytics. With greater attention to transparency in research design (i.e., reducing researcher degrees of freedom) and reporting (i.e., separating exploratory from confirmatory analyses), future studies can enhance systematic investigation into the constructs that facilitate multi-party co-located collaboration and provide us with essential foundation for robust automatic assessments.

CRediT authorship contribution statement

Y. H. Victoria Chua: Writing – review & editing, Writing – original draft, Visualization, Software, Investigation, Formal analysis, Data curation, Conceptualization. **Justin Dauwels:** Project administration, Funding acquisition, Conceptualization. **Preman Rajalingam:** Writing – review & editing, Resources, Project administration, Funding acquisition, Data curation. **Chew Lee Teo:** Supervision, Resources, Data curation. **Suzy J. Styles:** Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Conceptualization.

Acknowledgements

Work from this paper is supported by grants M4081917 and M4081918 from the Centre for Research and Development in Learning at NTU (CRADLE@NTU) and EdeX Teaching and Learning Grant M4082594.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2025.102215.

Data availability

The data utilised in this study is not publicly available due to ethical restrictions. The transcription protocol, processing scripts and analytical code is available at https://osf.io/y9btd/

References

Andrews-Todd, J., & Forsyth, C. M. (2020). Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Computers in Human Behavior, 104*, Article 105759.

Aronson, E. (1978). *The jigsaw classroom*. Sage.

Bachorowski, J. A., & Owren, M. J. (2001). Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect. *Psychological science, 12*(3), 252–257.

Bachour, K., Kaplan, F., & Dillenbourg, P. (2010). An interactive table for supporting participation balance in face-to-face collaborative learning. *IEEE Transactions on Learning Technologies, 3*(3), 203–213.

Bakhtiar, A., Webster, E. A., & Hadwin, A. F. (2018). Regulation and socio-emotional interactions in a positive and a negative group climate. *Metacognition and Learning, 13*, 57–90.

Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences*, 307–359.

Bassiou, N., Tsiartas, A., Smith, J., Bratt, H., Richey, C., Shriberg, E., … Alozie, N. (2016). *Privacy-preserving speech analytics for automatic assessment of student collaboration* (pp. 888–892). Interspeech.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of statistical software, 67*, 1–48.

Becheru, A., Calota, A., & Popescu, E. (2018). Analyzing students' collaboration patterns in a social learning environment using studentviz platform. *Smart Learning Environments, 5*(1), 1–18.

Berge, M. (2017). The role of humor in learning physics: A study of undergraduate students. *Research in Science Education, 47*, 427–450.

Berge, M., & Danielsson, A. T. (2013). Characterising learning interactions: A study of university students solving physics problems in groups. *Research in Science Education, 43*, 1177–1196.

Bergstrom, T., & Karahalios, K. (2007). Conversation Clock: Visualizing audio patterns in co-located groups. In *2007 40th annual Hawaii international conference on system Sciences (HICSS'07)*. IEEE, 78-78.

Borge, M., & Carroll, J. M. (2014). Verbal equity, cognitive specialization, and performance. In *Proceedings of the 2014 ACM international conference on supporting group work* (pp. 215–225).

Burgoon, J. K., Dunbar, N. E., & Giles, H. (2017). *Interaction coordination and adaptation*. Social Signal Processing.

Carmody, P. C., Mateo, J. C., Bowers, D., & McCloskey, M. J. (2017). Linguistic coordination as an unobtrusive, dynamic indicator of rapport, prosocial team processes, and performance in team communication. In *Proceedings of the human factors and ergonomics society annual meeting* (pp. 140–144). Sage CA: Los Angeles, CA: SAGE Publications, 61, No. 1.

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Chan, M. P. Y., Choe, J., Li, A., Chen, Y., Gao, X., & Holliday, N. R. (2022). Training and typological bias in ASR performance for world Englishes. *INTERSPEECH*, 1273–1277.

Chua, Y. H. V., Dauwels, J., & Tan, S. C. (2019). Technologies for automated analysis of co-located, real-life, physical learning spaces: Where are we now?. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 11–20).

Chua, Y. H. V., Rajalingam, P., Tan, S. C., & Dauwels, J. (2019). Edubrowser: A multimodal automated monitoring system for co-located collaborative learning. In *International Workshop on Learning Technology for Education in Cloud* (pp. 125–138). Cham: Springer International Publishing.

Chua, Y. H. V., & Styles, S. J. (2023). *The relationship of verbal contributions with learning gains and perceived teamwork quality in co-located collaborative learning [OSF Pre-Registered Report]*. In . https://doi.org/10.17605/OSF.IO/JB5A6.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association.

Claros, I., Cobos, R., & Collazos, C. A. (2015). An approach based on social network analysis applied to a collaborative learning experience. *IEEE Transactions on Learning Technologies, 9*(2), 190–195.

Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research, 64*(1), 1–35.

Cohen, E. G., & Lotan, R. A. (2014). *Designing groupwork: Strategies for the heterogeneous classroom third edition*. Teachers College Press.

Cornide-Reyes, H., Riquelme, F., Monsalves, D., Noel, R., Cechinel, C., Villarroel, R., … Munoz, R. (2020). A multimodal real-time feedback platform based on spoken interactions for remote active learning support. *Sensors, 20*(21), 6337.

Cress, U. (2008). The need for considering multilevel analysis in CSCL research—An appeal for the use of more advanced statistical methods. *International Journal of Computer-Supported Collaborative Learning, 3*(1), 69–84.

Cukurova, M., Zhou, Q., Spikol, D., & Landolfi, L. (2020). Modelling collaborative problem-solving competence with transparent learning analytics: Is video data enough?. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 270–275).

Dale, R., Fusaroli, R., Duran, N. D., & Richardson, D. C. (2013). The self-organization of human interaction. *Psychology of Learning and Motivation, 59*, 43–95.

Damon, W., & Phelps, E. (1989). Critical distinctions among three approaches to peer education. *International Journal of Educational Research, 13*(1), 9–19.

Deiglmayr, A., & Schalk, L. (2015). Weak versus strong knowledge interdependence: A comparison of two rationales for distributing information among learners in collaborative learning settings. *Learning and Instruction, 40*, 69–78.

Di Mitri, D., Schneider, J., Specht, M., & Drachsler, H. (2018). From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning, 34*(4), 338–349.

Dideriksen, C., Christiansen, M. H., Tylén, K., Dingemanse, M., & Fusaroli, R. (2023). Quantifying the interplay of conversational devices in building mutual understanding. *Journal of Experimental Psychology: General, 152*(3), 864.

Dillenbourg, P. (1999). What do you mean by" collaborative learning". *Collaborative learning-cognitive and computational approaches*.

DiMicco, J. M., Pandolfo, A., & Bender, W. (2004). Influencing group participation with a shared display. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work* (pp. 614–623).

Dowell, N. M., Nixon, T. M., & Graesser, A. C. (2019). Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions. *Behavior Research Methods, 51*, 1007–1041.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology, 23*(2), 283.

Dupont, S., Çakmak, H., Curran, W., Dutoit, T., Hofmann, J., McKeown, G., Pietquin, O., Platt, T., Ruch, W., & Urbain, J. (2016). Laughter research: A review of the ILHAIRE project. In A. Esposito, & L. C. Jain (Eds.), *Toward robotic socially believable behaving systems: Modeling emotions* (pp. 147–181). Springer International Publishing/ Springer Nature. https://doi.org/10.1007/978-3-319-31056-5_9.

Duran, N. D., Paige, A., & D'Mello, S. K. (2024). Multi-level linguistic alignment in a dynamic collaborative problem-solving task. *Cognitive Science, 48*(1), Article e13398.

Duran, N. D., Paxton, A., & Fusaroli, R. (2019). ALIGN: Analyzing linguistic interactions with generalizable techNiques—a Python library. *Psychological Methods, 24*(4), 419.

Ferguson, R., & Clow, D. (2017). Where is the evidence? A call to action for learning analytics. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 56–65).

Fusaroli, R., Rączaszek-Leonardi, J., & Tylén, K. (2014). Dialogue as interpersonal synergy. *New Ideas in Psychology, 32*, 147–157. https://doi.org/10.1016/j.newideapsych.2013.03.005

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive science, 40*(1), 145–171.

Gao, L., Li, X., Li, Y., & Hu, W. (2023). Capturing temporal and sequential patterns of socio-emotional interaction in high-and low-performing collaborative argumentation groups. *The Asia-Pacific Education Researcher, 32*(6), 817–831.

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Department of Statistics* (Vol. 348, pp. 1–17). Columbia University.

Gibson, C. B., Buchler, N., Hoffman, B., & La Fleur, C. G. (2019). Participation shifts explain degree distributions in a human communications network. *PLoS One, 14*(5), Article e0217240.

Gibson, D. R. (2003). Participation shifts: Order and differentiation in group conversation. *Social forces, 81*(4), 1335–1380.

Giles, H., & Gasiorek, J. (2013). Parameters of non-accommodation: Refining and elaborating communication accommodation theory. *Social cognition and communication*, 155–172.

Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2010). Language style matching as a predictor of social dynamics in small groups. *Communication Research, 37*(1), 3–19.

Goodwin, M. H., & Goodwin, C. (2012). Car talk: Integrating texts, bodies, and changing landscapes. *Semiotica, 2012*(191), 257–286.

Haan, K., Riedl, C., & Woolley, A. W. (2021). Discovering where we excel: How inclusive turn-taking in conversation improves team performance. In *Companion publication of the 2021 international conference on multimodal interaction (ICMI '21 companion), october 18-22, 2021, MontrÃ©al, QC, Canada* (p. 8). New York, NY, USA: ACM.

Haim, A., Shaw, S., & Heffernan, N. (2023). How to open science: A principle and reproducibility review of the learning analytics and knowledge conference. In *LAK23: 13th international learning analytics and knowledge conference* (pp. 156–164).

Hake, R. R. (2002). Assessment of student learning in introductory science courses. *PKAL Roundtable on the future: Assessment in the service of student learning,* (pp. 1–3). *Duke University*. March.

Haller, C. R., Gallagher, V. J., Weldon, T. L., & Felder, R. M. (2000). Dynamics of peer education in cooperative learning workgroups. *Journal of engineering education, 89*(3), 285–293.

Heuer, K., Müller-Frommeyer, L. C., & Kauffeld, S. (2020). Language matters: The double-edged role of linguistic style matching in work groups. *Small Group Research, 51*(2), 208–228.

Holler, J., Alday, P. M., Decuyper, C., Geiger, M., Kendrick, K. H., & Meyer, A. S. (2021). Competition reduces response times in multiparty conversation. *Frontiers in Psychology, 12*, Article 693124.

Holt, E. (2010). The last laugh: Shared laughter and topic termination. *Journal of Pragmatics, 42*(6), 1513–1525.

Hu, L., & Chen, G. (2021). Exploring turn-taking patterns during dialogic collaborative problem solving. *Instructional Science*, 1–26.

Janssen, J., Cress, U., Erkens, G., & Kirschner, P. A. (2013). Multilevel analysis for the analysis of collaborative learning. In *The international handbook of collaborative learning* (pp. 112–125). Routledge.

Jassawalla, A. R., Malshe, A., & Sashittal, H. (2008). Student perceptions of social loafing in undergraduate business classroom teams. *Decision Sciences Journal of Innovative Education, 6*(2), 403–426.

Jefferson, G. (1984). On the Organization of Laughter in Talk about Troubles. In J. M. Atkinson, & J. Heritage (Eds.), *Structures of Social Action: Studies in Conversation Analysis* (pp. 346–369). Cambridge: Cambridge University Press.

Kangasharju, H., & Nikko, T. (2009). Emotions in organizations: Joint laughter in workplace meetings. *Journal of Business Communication, 46*(1), 100–119, 1973.

Kapur, M. (2008). Productive failure. *Cognition and Instruction, 26*(3), 379–424.

Kauffeld, S., Lehmann-Willenbrock, N., & Meinecke, A. L. (2018). *21 the advanced interaction analysis for teams (act4teams) coding scheme*.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217.

Kim, B., & Ho, W. (2018). Emergent social practices of Singapore students: The role of laughter and humour in educational gameplay. *International Journal of Child-Computer Interaction, 16*, 85–99.

Kim, J., Truong, K. P., Charisi, V., Zaga, C., Lohse, M., Heylen, D., & Evers, V. (2015). Vocal turn-taking patterns in groups of children performing collaborative tasks: An exploratory study. In *16th annual conference of the international speech communication association, INTERSPEECH 2015* (pp. 1645–1649). International Speech Communication Association (ISCA).

Kitto, K., Manly, C. A., Ferguson, R., & Poquet, O. (2023). Towards more replicable content analysis for learning analytics. In *LAK23: 13th international learning analytics and knowledge conference* (pp. 303–314).

Koh, E., Lee, S. S., Caleon, I. S., Hong, H., Tan, J. P. L., Tee, Y. H., & Lek, H. H. (2020). *Measuring and nurturing teamwork competency through a computer-supported creative collaborative problem-solving programme*.

Koutsombogera, M., Costello, R., & Vogel, C. (2018). Quantifying dominance in the multisimo corpus. In *2018 9th IEEE international conference on cognitive infocommunications (CogInfoCom)* (pp. 141–146). IEEE.

Koutsombogera, M., & Vogel, C. (2019). Observing collaboration in small-group interaction. *Multimodal Technologies and Interaction, 3*(3), 45.

Koutsombogera, M., & Vogel, C. (2022). Understanding laughter in dialog. *Cognitive Computation, 14*(4), 1405–1420.

Kutnick, P., & Blatchford, P. (2014). Effective group work in primary school classrooms. Dordrecht: Springer Science & Business Media. Külekçi, E. ve Özgan, H.(2015). Üniversite Öğrencilerinin Algılarına Göre Sosyal Sorumluluk Alanında Görev Alma Nedenler ve Sonuçları, Anadolu Eğitim Liderliği ve Öğretim Dergisi, 3(2), 1-15.

Kyle, K.. Lexical_diversity. GitHub Repository. https://github.com/kristopherkyle/lexical_diversity.

Lai, C., Carletta, J., & Renals, S. (2013). Modelling participant affect in meetings with turn-taking features. In *Proc. Workshop of affective social speech signals*.

Lamminpää, J., & Vesterinen, V. M. (2018). The use of humour during a collaborative inquiry. *International Journal of Science Education, 40*(14), 1718–1735.

Leupen, S. M., Kephart, K. L., & Hodges, L. C. (2020). Factors influencing quality of team discussion: Discourse analysis in an undergraduate team-based learning biology course. *CBE-Life Sciences Education, 19*(1), ar7.

Lewis, C. M., & Shah, N. (2015). How equity and inequity can emerge in pair programming. In *Proceedings of the eleventh annual international conference on international computing education research* (pp. 41–50).

Linnenbrink-Garcia, L., Rogat, T. K., & Koskey, K. L. (2011). Affect and engagement during small group instruction. *Contemporary Educational Psychology, 36*(1), 13–24.

Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & d'Apollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research, 66*(4), 423–458.

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive science, 36*(8), 1404–1426.
Lu, L., Yuan, Y. C., & McLeod, P. L. (2012). Twenty-five years of hidden profiles in group decision making: A meta-analysis. *Personality and Social Psychology Review, 16*(1), 54–75.
MacNeil, S., Kiefer, K., Thompson, B., Takle, D., & Latulipe, C. (2019). Ineqdetect: A visual analytics system to detect conversational inequality and support reflection during active learning. In *Proceedings of the ACM conference on global computing education* (pp. 85–91).
Marshall, P., Hornecker, E., Morris, R., Dalton, N. S., & Rogers, Y. (2008). When the fingers do the talking: A study of group participation with varying constraints to a tabletop interface. In *2008 3rd IEEE international workshop on horizontal interactive human computer systems* (pp. 33–40). IEEE.
Martinez-Maldonado, R., Gašević, D., Echeverria, V., Fernandez Nieto, G., Swiecki, Z., & Buckingham Shum, S. (2021). What do you mean by collaboration analytics? A conceptual model. *Journal of Learning Analytics, 8*(1), 126–153.
McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods, 42*(2), 381–392.
Michaelsen, L. K., Parmelee, D. X., McMahon, K. K., & Levine, R. E. (2007). *Team-based learning for health professions education: A guide to using small groups*. Sterling, USA: Stylus Publishing LLC.
Motz, B., Brooks, C., Quick, J., Bergner, Y., Gray, G., Lang, C., Li, W., & Marmolejo-Ramos, F.. A baseline measure of open research practices in learning analytics. https://doi.org/10.35542/osf.io/325d7.
Nanyang Technological University. (2024). *FAQs - About Us. [Website]. Retrieved May 15, 2024*. from. In https://www.ntu.edu.sg/medicine/about-us/faqs#Content_C032_Col00.
Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606.
Oviatt, S., Hang, K., Zhou, J., & Chen, F. (2015). Spoken interruptions signal productive problem solving and domain expertise in mathematics. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (pp. 311–318).
Parmelee, D., Michaelsen, L. K., Cook, S., & Hudes, P. D. (2012). Team-based learning: A practical guide: AMEE guide no. 65. *Medical Teacher, 34*(5), e275–e287.
Perkins, D. N. (1993). Person-plus: A distributed view of thinking and learning. *Distributed cognitions: Psychological and educational considerations*.
Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences, 27*(2), 169–190.
Pickering, M. J., & Garrod, S. (2021). *Understanding dialogue: Language use and social interaction*. Cambridge University Press.
Platt, T., Hofmann, J., Ruch, W., & Proyer, R. T. (2013). Duchenne display responses towards sixteen enjoyable emotions: Individual differences between no and fear of being laughed at. *Motivation and Emotion, 37*, 776–786.
Poland, B., & Pederson, A. (1998). Reading between the lines: Interpreting silences in qualitative research. *Qualitative inquiry, 4*(2), 293–312.
Ponton, H., Osbourne, A., Greenwood, D., & Thompson, N. (2018). Understanding the dynamics of construction design team meetings through joint laughter. *Paper presented at 34th annual association of researchers in construction management conference, ARCOM 2018*. Belfast, United Kingdom.
Popov, V., van Leeuwen, A., & Buis, S. C. (2017). Are you with me or not? Temporal synchronicity and transactivity during CSCL. *Journal of Computer Assisted Learning, 33*(5), 424–442.
Praharaj, S., Scheffel, M., Drachsler, H., & Specht, M. (2021). Literature review on co-located collaboration modeling using multimodal learning analytics—can we go the whole nine yards? *IEEE Transactions on Learning Technologies, 14*(3), 367–385.
Radebe, N., & Mushayikwa, E. (2023). Bloom's Taxonomy and Classroom Talk: Exploring the Relationship Between the Nature of Small Group Discussion Tasks and the Quality of Learners. *Talk. African Journal of Research in Mathematics, Science and Technology Education*, 1–11.
Rajalingam, P., Rotgans, J. I., Zary, N., Ferenczi, M. A., Gagnon, P., & Low-Beer, N. (2018). Implementation of team-based learning on a large scale: Three factors to keep in mind. *Medical Teacher, 40*(6), 582–588.
Rasenberg, M., Özyürek, A., & Dingemanse, M. (2020). Alignment in multimodal interaction: An integrative framework. *Cognitive science, 44*(11), Article e12911.
Reilly, J. M., & Schneider, B. (2019). *Predicting the quality of collaborative problem solving through linguistic analysis of discourse* (pp. 149–157). International Educational Data Mining Society.
Remedios, L., Clarke, D., & Hawthorne, L. (2012). Learning to listen and listening to learn: One student's experience of small group collaborative learning. *Australian Educational Researcher, 39*, 333–348.
Reverdy, J., Koutsombogera, M., & Vogel, C. (2020). Linguistic repetition in three-party conversations. *Neural approaches to dynamics of signal exchanges* (pp. 359–370).
Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning* (pp. 69–97). Springer Berlin Heidelberg.
Russell, S. O. C., Gessinger, I., Krason, A., Vigliocco, G., & Harte, N. (2024). What automatic speech recognition can and cannot do for conversational speech transcription. *Research Methods in Applied Linguistics, 3*(3), Article 100163.
Salomon, G., & Globerson, T. (1989). When teams do not function the way they ought to. *International journal of Educational research, 13*(1), 89–99.
Shah, N., Lewis, C., & Caires, R. (2014). *Analyzing equity in collaborative learning situations: A comparative case study in elementary computer science*. Boulder, CO: International Society of the Learning Sciences.
Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal, 27*(3), 379–423.
Sinclair, A. J., & Schneider, B. (2021). *Linguistic and Gestural Coordination: Do Learners Converge in Collaborative Dialogue?* (pp. 431–438). International Educational Data Mining Society.
Spikol, D., Ruffaldi, E., & Cukurova, M. (2017a). *Using multimodal learning analytics to identify aspects of collaboration in project-based learning*. Philadelphia, PA: International Society of the Learning Sciences.
Spikol, D., Ruffaldi, E., Landolfi, L., & Cukurova, M. (2017, July). Estimation of success in collaborative learning based on multimodal learning analytics features. In *2017 IEEE 17th international conference on advanced learning technologies (ICALT)* (pp. 269–273). IEEE.
Stasser, G., & Titus, W. (2003). Hidden profiles: A brief history. *Psychological Inquiry, 14* (3/4), 304–313.
Stewart, A. E., Keirn, Z., & D'Mello, S. K. (2021). Multimodal modeling of collaborative problem-solving facets in triads. *User Modeling and User-Adapted Interaction*, 1–39.
Strid, E., & Cekaite, A. (2021). Embodiment in reciprocal laughter. *How emotions are made in talk, 321*, 163.
Strijbos, J. W., & De Laat, M. F. (2010). Developing the role concept for computer-supported collaborative learning: An explorative synthesis. *Computers in Human Behavior, 26*(4), 495–505.
Styles, S. J., Chua, Y. H. V., Woon, F. T., Liu, H., Garcia, L. P., Khudanpur, S., Khong, A. W. H., & Dauwels, J. (2023). Investigating model performance in language identification: Beyond simple error statistics. *Proceedings of INTERSPEECH 2023*. https://doi.org/10.21437/Interspeech.2023-1707
Subburaj, S. K., Stewart, A. E., Ramesh Rao, A., & D'Mello, S. K. (2020). Multimodal, multiparty modeling of collaborative problem solving performance. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (pp. 423–432).
Tausch, S., Hausen, D., Kosan, I., Raltchev, A., & Hussmann, H. (2014). Groupgarden: Supporting brainstorming through a metaphorical group mirror on table or wall. In *Proceedings of the 8th nordic conference on human-computer interaction: Fun, fast, foundational* (pp. 541–550).
Teng, C. C., & Luo, Y. P. (2015). Effects of perceived social loafing, social interdependence, and group affective tone on students' group learning performance. *The Asia-pacific education researcher, 24*, 259–269.
Trujillo, J. P. (2024-04-23). *Analyzing Multi-Party Turn-Taking Behavior*. Retrieved from:. In https://github.com/jptrujillo/turntaking_entropy_module/blob/main/TurnTaking_module.html.
Vogel, C., Koutsombogera, M., & Reverdy, J. (2023). Aspects of Dynamics in Dialogue Collaboration. *Electronics, 12*(10), 2210.
Vrzakova, H., Amon, M. J., Stewart, A., Duran, N. D., & D'Mello, S. K. (2020). March). Focused or stuck together: multimodal patterns reveal triads' performance in collaborative problem solving. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 295–304).
Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard university press.
Wang, L., Doucet, L., Waller, M., Sanders, K., & Phillips, S. (2016). A laughing matter: Patterns of laughter and the effectiveness of working dyads. *Organization Science, 27* (5), 1142–1160.
Watanabe, S., Mandel, M., Barker, J., Vincent, E., Arora, A., Chang, X., … Ryant, N. (2020). *CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings*. arXiv preprint arXiv:2004.09249.
Webb, B. R. (1995). Opinion: Educational research and computer supported Co-operative learning. *Innovations in Education and Training International, 32*(2), 139–146.
Weimar, E., Nugroho, A., Visser, J., Plaat, A., Goudbeek, M., & Schouten, A. P. (2017). *The influence of teamwork quality on software team performance. arXiv preprint arXiv:1701.06146*.
Wiltshire, T. J., Hudson, D., Belitsky, M., Lijdsman, P., Wever, S., & Atzmueller, M. (2021, September). *Examining team interaction using dynamic complexity and network visualizations* (pp. 1–6). IEEE.
Wise, A. F., Hsiao, Y. T., Marbouti, F., Speer, J., & Perera, N. (2012). Initial validation of "listening" behavior typologies for online discussions using microanalytic case studies. *Proceedings of the 10th International Conference of the Learning Sciences*, 56–63.
Wohltjen, S., & Wheatley, T. (2021). Eye contact marks the rise and fall of shared attention in conversation. *Proceedings of the National Academy of Sciences, 118*(37), Article e2106645118.
Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science, 330*(6004), 686–688.
Worsley, M., Anderson, K., Melo, N., & Jang, J. (2021). Designing analytics for collaboration literacy and student empowerment. *Journal of Learning Analytics, 8*(1), 30–34.
Xu, Y., & Reitter, D. (2017). Spectral analysis of information density in dialogue predicts collaborative task performance. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 623–633).
Yan, L., Zhao, L., Gašević, D., Li, X., & Martinez-Maldonado, R. (2023). Socio-spatial learning analytics in Co-located collaborative learning spaces: A systematic literature review. *Journal of Learning Analytics, 10*(3), 45–63.
Zambrano, J., Kirschner, F., Sweller, J., & Kirschner, P. A. (2019). Effects of prior knowledge on collaborative and individual learning. *Learning and Instruction, 63*, Article 101214.
Zhang, L., KaLyuga, S., Lee, C., & Lei, C. (2016). Effectiveness of collaborative learning of computer programming under different learning group formations according to students' prior knowledge: A cognitive load perspective. *Journal of Interactive Learning Research, 27*(2), 171–192.
Zhao, L., Yan, L., Gasevic, D., Dix, S., Jaggard, H., Wotherspoon, R., … Martinez-Maldonado, R. (2022, March). Modelling co-located team communication from voice

detection and positioning data in healthcare simulation. In *LAK22: 12th international learning analytics and knowledge conference* (pp. 370–380).

Zhou, N., Kisselburgh, L., Chandrasegaran, S., Badam, S. K., Elmqvist, N., & Ramani, K. (2020). Using social interaction trace data and context to predict collaboration quality and creative fluency in collaborative design learning environments. *International Journal of Human-Computer Studies, 136*, 102378.

Zhou, J., Hang, K., Oviatt, S., Yu, K., & Chen, F. (2014). Combining empirical and machine learning techniques to predict math expertise using pen signal features. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (pp. 29–36).

Zilvinskis, J. (2019). Measuring quality in high-impact practices. *Higher Education, 78*, 687–709.