Unveiling joint attention dynamics: Examining multimodal engagement in an immersive collaborative astronomy simulation

Jina Kang [a,*], Yiqiu Zhou [a], Robin Jephthah Rajarathinam [a], Yuanru Tan [b], David Williamson Shaffer [b]

[a] Curriculum and Instruction, University of Illinois Urbana-Champaign, Champaign, IL, USA
[b] Department of Educational Psychology, University of Wisconsin-Madison, Madison, WI, USA

ARTICLE INFO

Keywords:
Augmented and virtual reality
Cooperative/collaborative learning
Applications in subject areas
Simulations

ABSTRACT

Numerous computer-based collaborative learning environments have been developed to support collaborative problem-solving. Yet, understanding the complexity and dynamic nature of the collaboration process remains a challenge. This is particularly true in open-ended immersive learning environments, where students navigate both physical and virtual spaces, pursuing diverse paths to solve problems. In response, we aimed to unpack these complex collaborative learning processes by investigating 16 groups of college students ($n = 77$) who utilized an immersive astronomy simulation in their introductory astronomy course. Our specific focus is on joint attention as a multi-level indicator to index collaboration. To examine the interplay between joint attention and other multimodal traces (conceptual discussions and gestures) in students' interactions with peers and the simulation, we employed a multi-granular approach. This approach encompasses macro-level correlations, meso-level network trends, and micro-level qualitative insights from vignettes to capture nuances at different levels. Distinct multimodal engagement patterns emerged between low- and high-achieving groups, evolving over time across a series of tasks. Our findings contribute to the understanding of the notion of timely joint attention and emphasize the importance of individual exploration during the early stages of collaborative problem-solving, demonstrating its contribution to productive knowledge co-construction. This research overall provides valuable insights into the complexities of collaboration dynamics within and beyond digital space. The empirical evidence we present in our study lays a strong foundation for developing instructional designs aimed at fostering productive collaboration in immersive learning environments.

Funding

This material is based upon work supported by the United States National Science Foundation under Award 1822796. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

* Corresponding author.
E-mail addresses: jinakang@illinois.edu (J. Kang), yiqiuz3@illinois.edu (Y. Zhou), rjrthnm2@illinois.edu (R.J. Rajarathinam), yuanru.tan@wisc.edu (Y. Tan), dws@education.wisc.edu (D.W. Shaffer).

Code availability

The statistical code used to analyze the data during the current study is available from the corresponding author on reasonable request.

1. Introduction

Computer-supported collaborative learning (CSCL) environments facilitate students' co-construction of knowledge (Dillenbourg et al., 2009) and enhance collaborative problem-solving competencies (Sun et al., 2020). Immersive technologies, such as virtual, augmented, and extended reality offer unique affordances in facilitating collaborative learning around spatially complex science concepts (Dunleavy & Dede, 2014; Iba´n˜ez et al., 2014). These technologies allow learners to see holographic images of intangible scientific systems (micro to macro systems such as molecules to galaxies) in their physical world and use their hands to interact with the objects. Such potential of immersive technologies can be leveraged for learners to construct explanations for complex phenomena in astronomy that deeply involve spatial information including connecting between Earth-based and space-based reference frames (Albanese et al., 1997). Immersive technologies have been progressively integrated into collaborative learning environments. Various technology tools such as representational and relational tools are available to improve students' collaborative learning experience (Johri et al., 2013). Recent studies have explored the application of immersive technologies in STEM education, such as using AR and VR for teaching planetary science (Brenner et al., 2021) and cell biology (Webb et al., 2022), computer programming (Chung et al., 2021) and employing AR for circuits and electronics education (Villanueva et al., 2020).

As immersive technologies become more accessible as collaborative tools, they inherently reshape the collaboration dynamics where students use various modalities, including speech, eye gaze, and body movement, for communication and interaction. Advanced computational techniques (e.g., Cukurova et al., 2020) and collaboration analytics (Martinez-Maldonado et al., 2021; Wise et al., 2021) offer promising avenues for monitoring multimodal engagement and understanding learning in diverse technology-enhanced environments. However, challenges remain in analyzing the dynamic and complex nature of multimodal engagement in immersive settings.

Joint attention emerges as a critical phenomenon in collaborative learning. Joint attention facilitates shared meaning construction (Richardson et al., 2007) and promotes synergy among participants (Moore & Dunham, 1995; O'Madagain & Tomasello, 2021). The varying degrees of joint attention among students in a group can serve as a pivotal indicator of their coordination efforts during critical solution moments (Barron, 2000). Previous studies have explored joint attention using eye-tracking data and its relevance to collaboration quality (Schneider et al., 2018; Wisiecka et al., 2023), productivity (Liu et al., 2021), and learning gains (Abitino et al., 2022; Jermann et al., 2011). However, the complexity of immersive environments, with their open-ended three-dimensional virtual spaces, poses analytical challenges. Existing studies often rely on aggregated or binary measures, limiting their ability to capture nuanced collaboration dynamics (e.g., L¨amsa¨ et al., 2022).

To advance our understanding of students' collaborative learning in immersive learning environments, this study aims to delve deeper and comprehensively understand the implications of various degrees of joint attention by integrating different modalities, logs, discourse, and gestures. This enhances our understanding of how joint attention contributes to collaborative knowledge construction. We employ an innovative analytical approach to unravel how students work together through physical and virtual spaces and how their collaborative behaviors are associated with their learning performance. As such, this study bridges gaps in the existing literature by employing a novel approach that enables a finer-grained exploration of joint attention within real-world settings such as classrooms. This approach particularly accounts for varying levels of students' multimodal engagement and transcends the binary categorization of joint attention moments. By unpacking the interplay between joint attention and conceptual engagement, we aim to depict interaction portraits that capture the dynamics in groups' sense-making over time and elucidate the variations in collaborative learning outcomes. We propose three research questions as follows:

- RQ1: What relationships exist between joint attention and learning performance during CPS in an immersive astronomy simulation?
- RQ2: In what ways joint attention, gesture, and turn-taking (speech) work together to enhance the comprehension of group conceptual discussions?
- RQ3: How does the interplay between joint attention and conceptual discussions evolve over time among groups with varying learning performances?

The following section outlines the theoretical framework. We begin by introducing joint attention as a key construct for understanding collaborative learning and the need for new methods to trace the dynamics of joint attention. Subsequently, we delve into the necessity for innovative methods to further unpack collaborative learning processes and introduce our analytical approach, Ordered Network Analysis (ONA).

2. Theoretical framework

2.1. Joint attention

Joint attention refers to the ability to coordinate attention toward a partner or an object of mutual interest (Bakeman & Adamson, 1984). From both a social-cognitive lens (Moore & Dunham, 1995; O'Madagain & Tomasello, 2021) and a socio-constructivist perspective (Palincsar, 1998), joint attention plays a pivotal role in collaboration. It is instrumental for communication and effective collaborative problem-solving, as it promotes shared meaning construction (Richardson et al., 2007), helps establish a shared problem-solving space, and promotes synergy among participants (Moore & Dunham, 1995; O'Madagain & Tomasello, 2021). Barron (2000) perceived the degree of joint attention during solution-critical moments as a key dimension of coordination, with the depth and intensity of such attention potentially explaining variations in learning performance.

Given this theoretical emphasis on the role of joint attention in collaboration, numerous empirical studies have employed eye-tracking devices to measure joint visual attention, a specific facet of joint attention characterized by the mutual coordination of eye gaze direction (Schneider & Pea, 2013; Sharma et al., 2017). Deriving metrics from eye movement data, including gaze similarity or cross-recurrence analysis for gaze convergence (Jermann & Nüssli, 2012; Richardson & Dale, 2005), these studies offer substantial evidence for a positive correlation between joint visual attention and collaborative learning outcomes. For instance, high levels of joint attention may indicate collective cognitive responsibility, where group members actively monitor and adapt to each other's focus. This collective responsibility can enhance the productivity of dyads (Liu et al., 2021) by ensuring that individual contributions align with shared goals. By jointly focusing on key elements together, students can reduce cognitive load and enhance the overall collaboration quality (Schneider et al., 2016, 2018; Wisiecka et al., 2023). Moreover, high gaze cross-recurrence may reflect cumulative efforts to establish mutual understanding, which can lead to improved learning and task performance (Abitino et al., 2022; Jermann et al., 2011).

Traditional eye-tracking devices, however, have limitations due to environmental and calibration requirements (e.g., precise camera positioning) and their incapability to track three-dimensional projections or measure attention quality (Bovo et al., 2022). These constraints are particularly evident in open-ended virtual environments, where students freely navigate a three-dimensional world and shift between scenes instead of being confined to a two-dimensional interface. Furthermore, existing methods primarily assess the presence of joint visual attention as an indicator of collaborative moments (Schneider & Bryant, 2022). However, this approach is insufficient in comprehending sophisticated collaborative behaviors, as it often quantifies joint activity through an aggregated proportionate representation of time spent on observing the same object (L¨ams¨a et al., 2022). This simplified representation can obscure unproductive scenarios, such as the free-rider effect (Schneider et al., 2018) or misplaced attention on irrelevant objects (Hahn & Klein, 2022). The literature presents mixed findings on the relationship between joint attention and learning outcomes, suggesting that a binary representation is overly simplistic and potentially misleading.

To overcome these limitations, this study explores the temporal dynamics of joint attention. We expand on our previous method (Diederich et al., 2021), which involves measuring joint attention by monitoring the degree of screen overlap across different devices. Joint attention levels are computed by tracking both user head movement and the field of view within the simulation. This allows us to examine joint attention without impeding students' device usage, eliminating the need for calibration as required in eye-tracking devices. Such visual synchronization can require varying levels of effort, particularly when students have to adjust their headsets or screens to locate specific objects. It reflects a conscious decision to establish joint attention through intentional communication and awareness of each other's focus (Siposova & Carpenter, 2019). Our preliminary study (Zhou & Kang, 2022) yielded promising results and intriguing patterns of joint attention within an immersive learning environment. This current study aims to illustrate the significance of exploring the temporal aspects of joint attention and its importance as a key construct for gaining deeper insights into collaborative learning.

2.2. Learning analytics in computer-supported collaborative learning

Prior research on collaborative learning has mainly relied on the coding of video data (Hmelo-Silver & Barrows, 2008), which is time-consuming, particularly for large-scale interventions. CSCL environments offer opportunities to trace individual students' actions and their interactions with peers within the environment. To complement qualitative coding, recent research has introduced data-driven approaches using various features extracted from digital traces. Most studies developed uni-modal features to predict students' performance or build students' behavior models. However, such approaches are still limited to unraveling dynamic and interactive collaborative processes within CSCL environments and accurately interpreting and accounting for factors influencing learning (Cukurova et al., 2020). For instance, unstructured logs generated from open-ended environments pose challenges to understanding how students navigate the environment and work with their peers and further assess how certain interaction behaviors are associated with positive learning opportunities (Akçayır & Akçayır, 2017).

To advance our understanding of students' collaborative learning in CSCL environments, recent studies have explored the potential of multimodal data generated from multiple sensors (Di Mitri et al., 2019; Sharma & Giannakos, 2020). The term "multimodal data" refers to forms of data that are not limited to written or spoken language (Scollon & Scollon, 2009), including linguistic, behavioral, embodied, spatial, visual, and physiological aspects of communication and meaning-making. According to Vrzakova et al. (2020), verbal and nonverbal group behaviors are significantly associated with meaningful learning outcomes. They demonstrated the advantages of the multimodal features to unveil the dynamics of collaborative problem-solving processes and highlighted the need to understand the contribution of each modality. As such, the field of multimodal learning analytics (MMLA) emerged to provide insights beyond traditional analytics (Giannakos et al., 2022; Martinez-Maldonado et al., 2019). MMLA can provide a more holistic picture of learning processes and success factors. It uses the advances in machine learning and sensor technologies to monitor factors that are argued to be significant for learning but are often ignored due to challenges in their dynamic measurement and interpretations (Cukurova et al., 2020).

Temporality is another essential aspect for gaining a comprehensive view of collaboration. Process-oriented research focuses on the collaborative process rather than the collaborative outcomes (Dillenbourg, 1999; Janssen et al., 2010). Capturing the diverse collaborative processes and their outcomes during a collaborative experience is vital to ensuring successful collaboration in CSCL contexts (Ludvigsen et al., 2018). Such approaches have been applied in a more traditional sense to gain a comprehensive understanding of why some groups fail (e.g., Barron, 2003; Van de Pol et al., 2019). Recently, various learning analytics techniques such as process mining have been applied to investigate sequential and temporal characteristics of learning processes captured by digital traces (e.g., Saint et al., 2022). Different network analytic approaches such as Epistemic Network Analysis (ENA) were applied to examine the sequential and temporal nature of collaborative learning behaviors (e.g., Melzner et al., 2019). Researchers have introduced Ordered Network Analysis (ONA), as an extension of ENA, to model collaborative learning where the temporal order of events is hypothesized to be meaningful. Previous studies (Tan et al., 2022, Fan et al., 2023) have emphasized the analytical benefits of ONA in advancing our understanding of learning, particularly in analyzing log data, collaborative discourse data, and spatial position data.

First, as a widely used network technique for modeling learning phenomena (Shaffer, 2017; Swiecki et al., 2020), ONA builds on the strengths of ENA, such as modeling collaborative learning as interconnected interactions among individuals over time rather than a set of isolated events. Additionally, ONA considers the sequence of interactions in these processes in both its modeling and visualization algorithms. Specifically, in its modeling phase, ONA constructs asymmetrical connection matrices to record both the connection strength and connection directions between any pairs of actions. In its visualization phase, ONA produces directed networks that can be used to interpret different groups' connection patterns in different stages. As a result, it is particularly beneficial when research aims to understand the evolution of the interplay between different features over time. To better understand the behavior of specific groups, ONA considers the recent temporal context in which the behavior occurred and models their behavior accordingly. It then creates a model for all groups' networks using a set of deterministic node positions. This approach enables a fair comparison of how groups behave differently under the same experimental settings. Lastly, the ONA R package (Marquart et al., 2022) offers various advanced mathematical tools, such as dimensional reduction and co-registration, which ensure that the ONA networks produced can be easily interpreted and compared visually, as well as subjected to statistical analysis. This study, therefore, seeks to enhance comprehension of the processes involved in a series of students' collaborative problem-solving tasks by investigating both multimodality and temporality.

3. Methods

3.1. Participants

The participants included 77 undergraduates enrolled in an introductory astronomy course from a mid-western university in the United States. This course is designed for non-majors to fulfill general education requirements, with one main lecture and seven smaller discussion sections. To ensure diverse representation, participants were randomly selected from three of these seven discussion sections. During the study, students participated in over three weekly 50-min lab sessions: 2 introductory and 1 immersive learning simulation sessions. The simulation (CEASAR: Connections of Earth And Sky with Augmented Reality) allows for the exploration of the night sky from three different scenes (Horizon, Star, Earth; see Fig. 1 and Appendix A). They were expected to utilize the scenes to solve

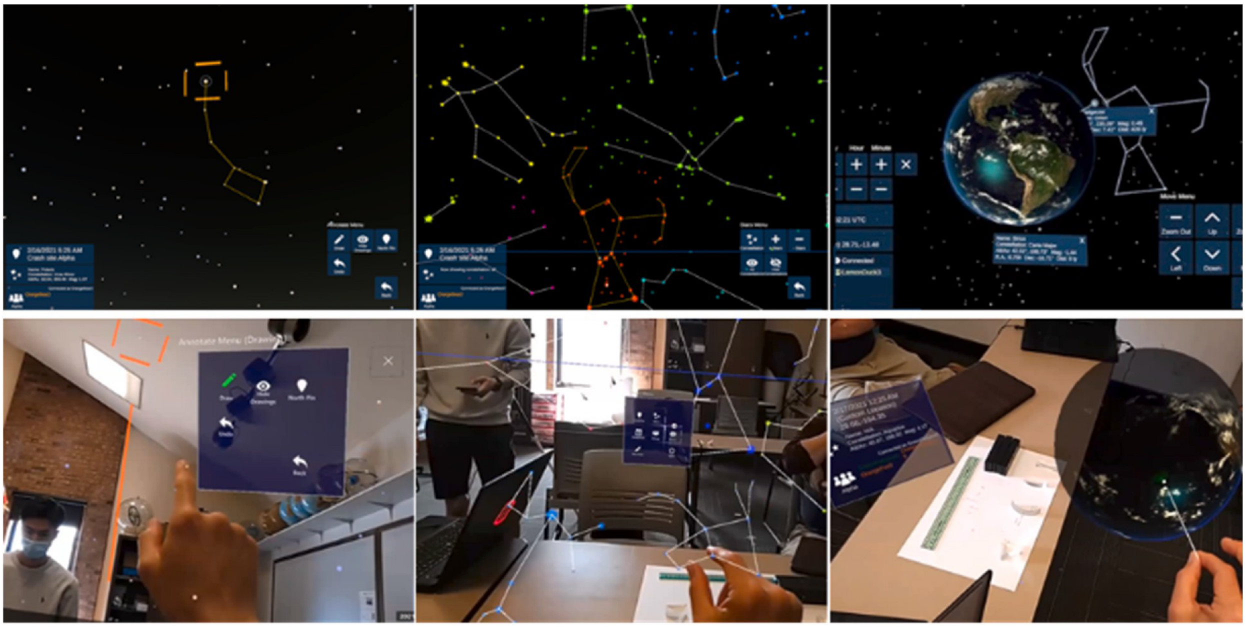


Fig. 1. Screenshots of CEASAR tablet (top) and AR views (bottom): Horizon (left), Start (mid), and Earth scene (right).

a group task, Lost at Sea (see details in 3.2). Since this simulation was designed to support collaboration, it synchronizes users' input (e.g., annotations, highlighted stars), and location across devices within the group. A group of 3–4 students was given one AR headset and two touch-based tablets. The first two sessions provided an introduction to using simulations and AR headsets. Students engaged in the group task during the last session. The participants chose which groups to join, resulting in a total of 25 groups. Participants had the option to work at a table where no data was collected. 16 groups remained for further analyses, accordingly.

3.2. *Tasks*

The participants were engaged in a multi-part problem-solving task called "Lost at Sea." In this scenario, a crewed space capsule splashed down in an unknown ocean location on Earth at night (see Appendix B). To determine the approximate latitude and longitude of their capsule splash-down site, the students used the simulation. The problem is divided into four parts. Firstly, each group identifies stars or reference constellations to determine the hemisphere where the capsule was located (Task-1). Task-2 requires the group to identify constellations that serve as reference points for North, South, East, and West for the crashed crew. The last two tasks involve formally calculating the latitude (Task-3) and longitude (Task-4) of the capsule's splash-down site, respectively.

3.3. *Data sources*

This study particularly used the data collected during the simulation session. We leveraged multiple data sources, including audio/video recordings, log files, screen captures from devices, and assessment data. Our main goal was to capture both verbal and nonverbal interactions related to joint attention (Jermann & Nüssli, 2012; Sharma et al., 2013). By using multimodal data, we aimed to provide a comprehensive account of how students work together to solve a group task using the simulation. Log files were generated as students interacted with the simulation, which allowed us to identify different levels of JA of the group at the second granularity (see Table 1; Diederich et al., 2021).

Our research team developed a coding scheme (Planey et al., 2023), focusing on verbal and non-verbal cognitive interactions captured in audio and video recordings (see Table 1). The verbal codes documented specific task-relevant discussions when students collaborate on open-ended tasks, using turn-by-turn coding (Mercier et al., 2017; Shehab & Mercier, 2020). Gestures were coded as non-verbal conceptual interactions that help students discuss complex astronomy phenomena. Two researchers manually coded one group's data independently, addressing any discrepancies through reconciliation. Subsequently, 20 % of the remaining data underwent manual coding and double-checking for agreement, ensuring an inter-rater reliability of at least 80 % across all categories. Following this initial validation, researchers individually coded the remaining dataset. Each data was synchronized at the same second for further analysis.

Lastly, we collected pre-/post-assessments to measure individuals' conceptual knowledge. In this study, we focused on an open-ended question to evaluate students' understanding of latitude and longitude calculation: "*Write as much as you know about the steps for calculating the latitude and longitude based on the stars visible in a given location.*" Individual responses were scored from 0 to 2 based on completeness and accuracy: 0 for incorrect answers, 1 for partially correct, and 2 for full understanding of the concept.

3.4. *Features*

The **simulation interaction features** include five JA states, indicating varying levels of joint attention. As shown in Table 1, these states range from inactivity (*JA I*) to various levels of scene engagement (*JA II* and *III*) and screen overlap (*JA IV* and *V*). Specifically, *JA IV* and *V* were determined by assessing the overlap in field-of-view across different devices, informed by user head movements and in-simulation viewpoints (Zhou & Kang, 2022). The initial JA features were generated for three device pairs per group (Tablet1-AR, Tablet1-Tablet2, Tablet2-AR). The pair with the highest level of JA was selected to represent the whole group's joint attention behavior.

The **conceptual interaction features** include (1) verbal: *conceptual turn-taking* and four conceptual discussion codes, including *new knowledge*, *modify knowledge*, and *confirm knowledge*, and *confusion*, and (2) non-verbal: *gesture*. *Conceptual turn-taking*, determined by the total number of task-relevant discussion rounds made by various group members within a single episode, reflects a group's

Table 1
Feature descriptions.

Feature	Description
JA I: Inactivity	No events triggered within the 20 s
JA II: No scene overlapping	Students explore in different scenes OR only one student is active
JA III: Scene overlapping in Earth or Star	Both students stay in Earth or Star scene
JA IV: No scene overlapping in Horizon	Both students stay in the Horizon scene without any screen overlaps
JA V: Scene overlapping in Horizon	A screen overlap in the Horizon scene was observed across two devices
New knowledge [NK]	Students introduce new information or concepts.
Modify knowledge [MK]	Students modify or build on existing knowledge.
Confirm knowledge [CK]	Students support or restate existing knowledge.
Confusion [CO]	Students are unsure of how to interpret information or task requirements.
Gesture	Physical motions representing concepts or directing attention.

active engagement in the discussion.

All features were aggregated by an **episode**, ending with a 10-s pause in task-relevant discussion. This 10-s threshold was determined to capture related exchanges while minimizing coder inconsistency (Planey et al., 2023). Specifically, JA states were aggregated based on duration, while all other features were based on the count of occurrences within an episode. We then weighted each feature by the episode length to ensure standardization across all episodes.

Lastly, *normalized learning gains* (learning performance) were calculated from the pre-/post-tests: (post−pre)/((post_max)−pre) (Hake, 1998). This approach normalizes the learning gains relative to the maximum possible improvement, thereby allowing for a fair comparison across varying initial knowledge levels. Upon analyzing the distribution of group average gains, a noticeable gap emerged between 0.1667 and 0.3125. This gap indicates a potential divergence in knowledge acquisition, suggesting groups either learned less than 16.67 % or more than 31.25 % of the maximum possible gains. We subsequently categorized 16 groups into low-achieving ($n = 7$) with a range of [−0.2222, 0.1667] and high-achieving groups ($n = 9$) with a range of [0.3125, 0.6875].

3.5. Analyses

Our study performed analysis across three levels: all-tasks level (aggregating data across all four tasks), task level (aggregating data for each task), and episode level (aggregating data for each conversation episode). Each offers unique insights. To address RQ1, we identified differences between low and high-achieving groups at the all-tasks level. After evaluating the normality of feature distribution, we employed Welch's *t*-test and a two-sample Wilcoxon test respectively for the corresponding features. This allows for the comparison of time spent on each task and on each JA state across groups with varying learning performances.

To address RQ2, we integrated three unimodal features (*conceptual turn-taking*, five JA states, and *gesture*) to generate bi/multimodal interaction features (see examples in Appendix F). *Conceptual turn-taking* and *gesture* are essential components of communication and knowledge construction (Schneider et al., 2021). Incorporating bi/multimodal features can lead to a more comprehensive understanding of collaboration behaviors. Pearson correlation analyses were performed at the episode level to examine the associations between the uni/bi/multimodal features and conceptual discussion codes. While such analyses examine various combinations of data streams, the interplay between these multimodal features that contribute to productive interactions remains unclear.

In response, RQ3 sought to provide a more comprehensive insight into the collaborative learning processes by examining the interplay of multimodal features. To do so, we initially selected vignettes that best-exemplified collaboration processes, with a specific focus on the last two tasks due to their demanding nature, involving complex mathematical computations. We examined both high- and low-achieving groups separately to observe how students build mutual agreement (*confirm knowledge*) and adopt new ideas from their group mates (*new knowledge*). We examined the entire coded datasets, videos, and screen captures to identify vignettes that met these conditions. The conceptual interaction codes were first filtered for the occurrence of *new* and *confirm knowledge* and then used to identify sequences of related conceptual interaction codes. The filtered results were then used to identify two exemplar vignettes for each of the high- and low-achieving groups, specifically due to their availability to demonstrate diverse multimodal engagement strategies employed by the groups.

Building on these qualitative insights, we leveraged ONA as network analysis to identify emerging patterns of interactions across modalities. While the narrative-driven vignettes served as snapshots of collaboration instances, the ONA revealed the connections from a more quantifiable lens, identifying patterns that might not be immediately apparent in the vignettes. This approach not only deepened our insights into feature associations but also uncovered mechanisms underlying successful collaborative learning, drawn from vignettes, thereby *closing the interpretive loop* (Shaffer, 2017).

Specifically, we employed the ONA R package (Marquart et al., 2022) to conduct network analysis at each task level. The model parameters included *codes*, *units of analysis*, *conversations*, and *moving stanza window size*. In our study, *codes* consist of five simulation interaction features and four conceptual discussion features described in Table 1. Based on the results from RQ2, we removed gestures as they were limited to offer a clear advantage over bimodal features with conceptual turn-taking, aiming to balance between information depth and visual simplicity. For *units of analysis,* we used the unique combinations of "group", "task", and "performance." This means that groups will be compared based on their performance categories (i.e., low and high-achieving) and tasks, regarding how they made connections with the 9 *codes*. This guided our specification for *conversation*; we used the unique combinations of "performance" and "task" assuming no cross-group interactions during tasks, which aligned with our classroom observations. Lastly, we used a *moving stanza window* size of four lines based on statistical optimization and qualitative evaluations. The 4-turn time window size resulted in the highest model fit and largest variance explained (see Appendix C) after testing multiple configurations (Ruis et al., 2019). Additionally, it aligns with the natural rhythm and flow of the conversations, as confirmed by examining interaction vignettes and dialog coherence. We thus selected this window size to highlight pronounced differences between high and low-achieving groups. Once the model parameters have been specified, the ONA algorithm tracked directed connections within the *moving stanza window* for each *unit of analysis*. Connections were aggregated across *conversations* and represented as high-dimensional vectors. ONA leverages a dimension reduction technique via the means rotation (MR; Bowman et al., 2021; Fan et al., 2023) method to project high-dimensional vectors into a two-dimensional space. In this study, MR is used to maximize the differences between the high- and low-achieving groups. The network nodes, which represent the codes defined earlier, were positioned within this space through an optimization routine coregistration. This setup enables interpretation based on nodes and locations of analysis units. For example, units primarily connecting to nodes on the left side of the space will be positioned to the left, and those with comparable connection patterns will be located closer to each other. Each *unit of analysis* was visualized in the resulting ONA networks with two distinct representations: (1) an ONA point, representing the location of its network in the two-dimensional projected space, and (2) a directed weighted network where nodes correspond to the codes, and edges reflect the relative frequency and direction of connection between two codes. We also conducted Welch's two-sample *t*-test on the distribution of the ONA points to determine if the differences between the two performance categories were statistically significant.

4. Results

4.1. RQ1 what relationships exist between JA and learning performance during CPS in an immersive astronomy simulation?

We first investigated whether the duration to complete each task is different between low and high-achieving groups (see Appendix D). We analyzed each task separately as the range of coordinated actions and required knowledge to complete the task were different. High-achieving groups demonstrated a tendency to allocate less time to the first two tasks while devoting more time to the subsequent two tasks, an opposite trend observed in low-achieving groups. This "positive time on task" effect for high-achieving groups could possibly reflect a strategic allocation of time and cognitive resources among high-achieving groups (Goldhammer et al., 2014). However, Welch's *t*-test failed to reveal any significant difference between low- and high-achieving groups concerning the time devoted to each task. We further examined the cumulative time spent on each JA state in order to see any relationship between group-level attention coordination behavior (i.e., the average amount of time spent on each JA state) and their learning gains. The results (see Appendix E) showed that high-achieving groups were more engaged in higher levels of JA (specifically *JA IV* and *JA V*) throughout the entire problem-solving process. However, the comparison tests did not identify any significant difference in the duration between high- and low-achieving groups. These non-significant differences might be tied to insufficient statistical power, a challenge often encountered in collaborative learning research due to a small sample size (Manathunga & Herna´ndez-Leo, 2015). This also highlights the limitation of a static measure of the group-level construct–derived by aggregating the durations of JA states over time–that is insufficient to accurately capture the complexity of JA dynamics.

4.2. RQ2 in what ways do JA, gesture, and turn-taking work together to contribute to understanding group conceptual discussions?

We further conducted a Pearson correlation analysis at the episode level to investigate how simulation interaction features (JA states) combined with conceptual interaction features (*conceptual turn-taking* and *gesture*) contribute to group sense-making and knowledge development (conceptual discussion codes). We initiated our analysis by examining whether the unimodal feature, each JA state, alone could offer insights into the knowledge co-construction. Notably, different JA states correlated uniquely with conceptual

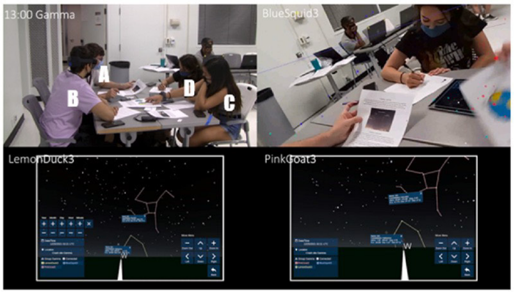
A: Because I know it has something to do with the angle of the stars [NK] … Here finding latitude (picks up paper)

B(AR): There we go [CK]

A: (Reads supplementary material aloud) [NK]

B(AR): We got to find the angle? [CK]

A: Go to Polaris [NK] and find the angle, it's the latitude [NK]



A: At what point… At what angle is this?

C(AR): Polaris is supposed to be on this- [NK]

B: The second cap- you have to look up the constellation ursa minor [NK]

A: Oh yeah [CK]. Is that the one that's supposed to be-

B: You have to go to constellation, click on there, and it will show you where it is

C(AR): I think I dropped a pin that I think is fairly close I think [NK]

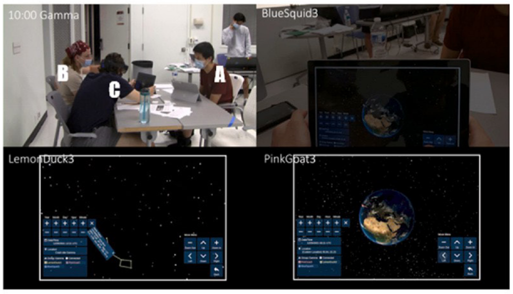


Fig. 2. Vignettes of a high (top) and low-achieving group (bottom) during Task-3. Each vignette includes transcription dialog (left), group interaction (top-right), and screens (top-bottom).

discussion codes, suggesting that different types of JA may play distinct roles in the group sense-making process (see Appendix F). For instance, *JA IV* was positively correlated with new knowledge ($r(184) = 0.161$ $p < 0.05$)and negatively correlated with confusion ($r(184) = -0.186$ $p < 0.05$). Given that *JA IV* occurs when both devices interact with the Horizon scene without screen overlap, it signifies an individual exploration stage where each device explores different areas of the simulated sky. The finding showed this state potentially contributes to the discovery of *new knowledge* and reduces *confusion*. Additionally, *JA V*, characterized by screen overlapping, exhibited a positive association with *confirm knowledge* ($r(184) = 0.291$ $p < 0.001$). This connection implies that screen overlapping helps create a shared problem space to ground group discussion (Stahl, 2006).

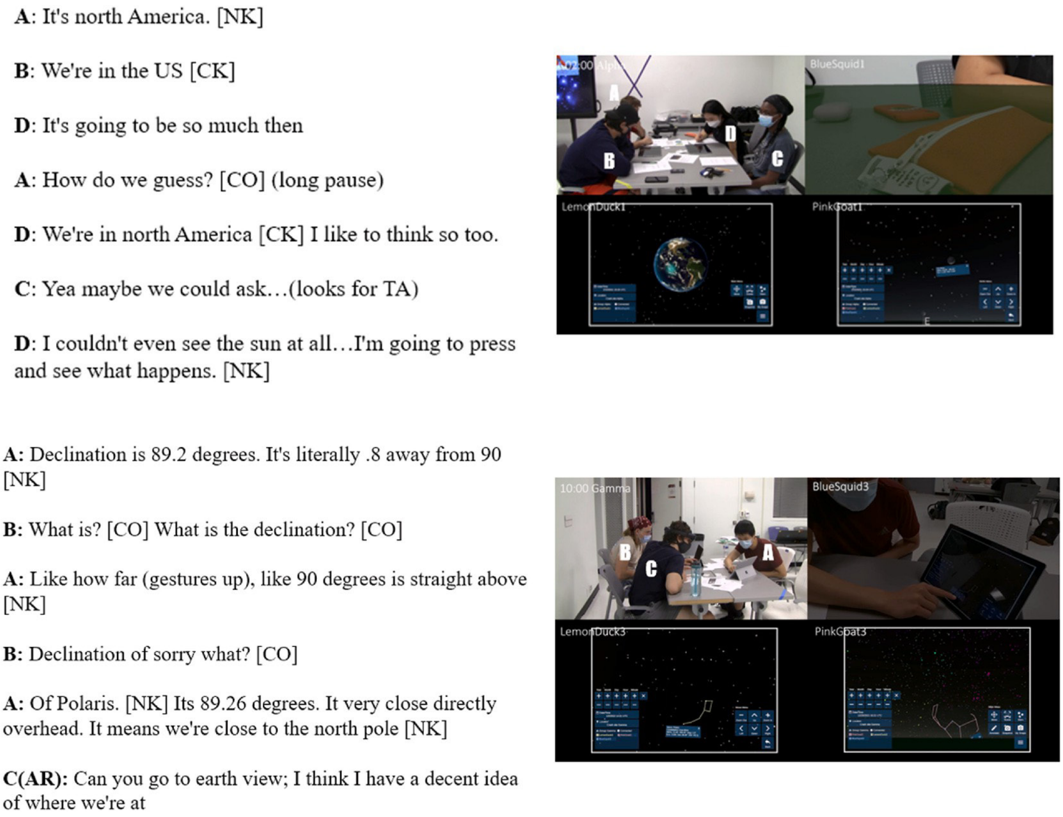
Building upon these findings, we delved deeper into the advantages of incorporating conceptual interaction features as additional modalities. To assess this added value, we compared the correlations between unimodal and bi/multimodal features. We employed Zou's test, specifically tailored for discerning differences between two overlapping dependent correlations with one common variable (i.e., each JA state) (Zou, 2007). This comparison determines whether additional modalities significantly increased the correlations with conceptual discussion codes. Our results (see Appendix F) suggested that incorporating *conceptual turn-taking* enhanced the magnitude of the correlation significantly. For example, the combination of *conceptual turn-taking* and *JA IV* yields a stronger positive correlation with *new knowledge*. Adding gestures led to marginal decreases or similar coefficients in correlation strength in the context of *new knowledge, modify,* and *confirm knowledge*. These findings together suggest that gestures did not consistently bring notable advantages over the bimodal features with *conceptual turn-taking*. The role of gestures seems to be more intricate and warrants further exploration of its types and context.

4.3. RQ3 how does the interplay between joint attention and conceptual discussions evolve over time among groups with varying learning performances?

4.3.1. Vignettes of interaction

To delve deeper into the interplay between joint attention and conceptual discussions, we identified vignettes that best showcased knowledge-building processes like shared understanding and adoption of new ideas within each performance group during Task-3 and Task-4.

Vignette 1 (Task-3, High-achieving group) shows four students collaboratively determining the crash site's latitude (Fig. 2, top). Student A, holding the paper resource, aided Students C and D in identifying the constellation Polaris. This resulted in joint attention as they focused on the scene on their tablets. Although initially uncertain about how to determine latitude, Student A later shared *new knowledge*, emphasizing the importance of star angle [NK] and paper resources to determine latitude. Student B affirmed Student A's



A: It's north America. [NK]

B: We're in the US [CK]

D: It's going to be so much then

A: How do we guess? [CO] (long pause)

D: We're in north America [CK] I like to think so too.

C: Yea maybe we could ask…(looks for TA)

D: I couldn't even see the sun at all…I'm going to press and see what happens. [NK]

A: Declination is 89.2 degrees. It's literally .8 away from 90 [NK]

B: What is? [CO] What is the declination? [CO]

A: Like how far (gestures up), like 90 degrees is straight above [NK]

B: Declination of sorry what? [CO]

A: Of Polaris. [NK] Its 89.26 degrees. It very close directly overhead. It means we're close to the north pole [NK]

C(AR): Can you go to earth view; I think I have a decent idea of where we're at

Fig. 3. Vignettes of a high (top) and low-achieving group (bottom) during Task-4.

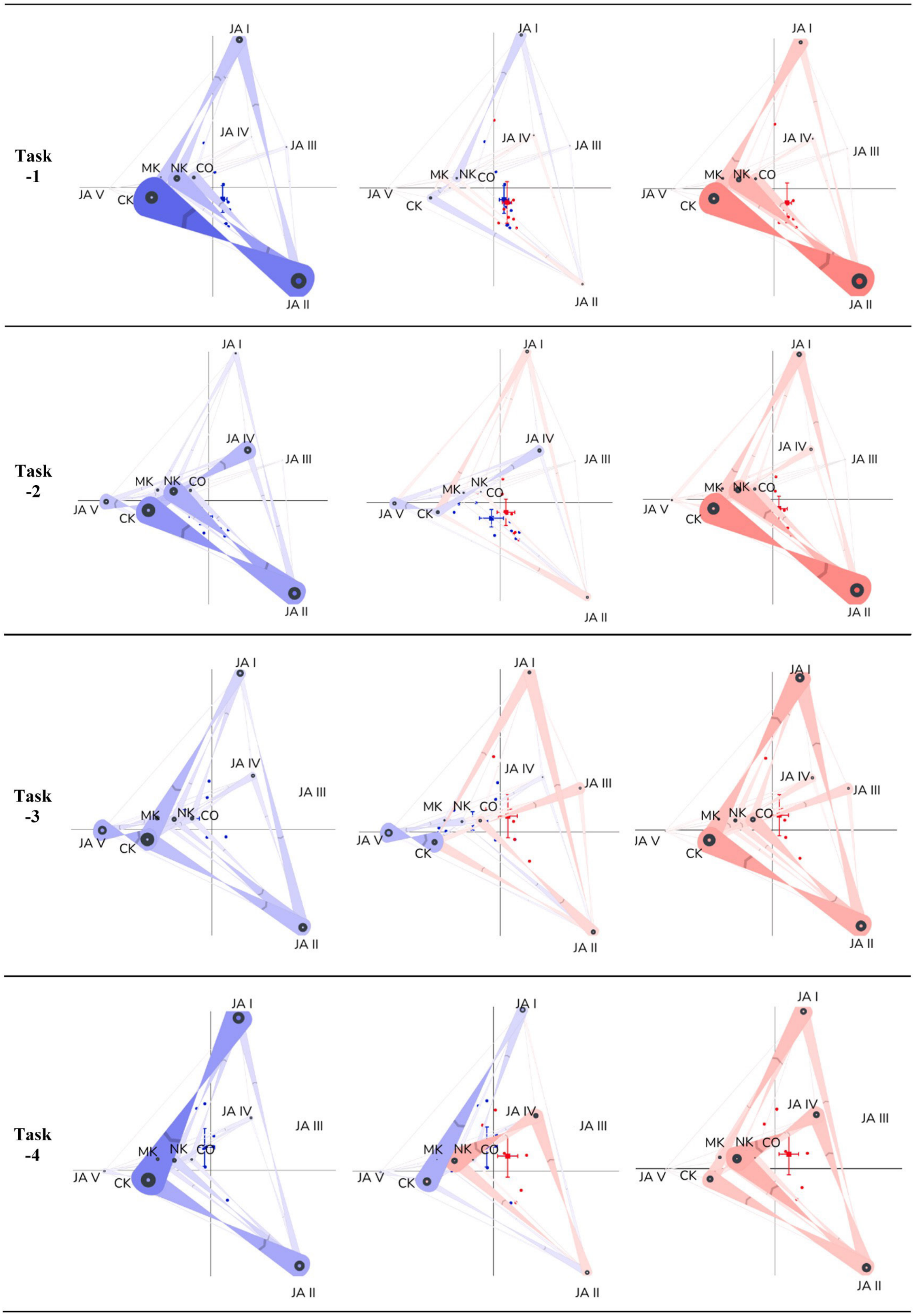


Fig. 4. ONA networks of high- and low-achieving groups' multimodal engagement patterns for each task. High-achieving groups (blue; left), a subtracted network (mid), and low-achieving groups (red; right). A node represents each feature. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

actions [CK] and listened to the explanation. Student B then sought and received confirmation from Student A [CK] who consulted the resource [NK]. This exchange facilitated shared knowledge co-construction, as the students built on initial ideas, learned from one another, and successfully determined their latitude.

Vignette 2 (Task-3, Low-achieving group) reveals the lack of overlapping views and mutual engagement in the same virtual scene. This lack of joint attention was mirrored in their challenges to reach a consensus with students often working independently and focusing on their own screens (Fig. 2, bottom). For instance, while Student A and B interacted, Student C was engrossed in searching for Polaris in the Earth scene and ignored Student A's question about the constellation angle. Student B responded to A's question with insights on Ursa Minor [NK]. Despite confirming this information [CK], Student A still faced difficulties locating the constellation. This group's focus on separate tasks and lack of consensus hindered their effective collaboration.

Vignette 3 (Task-4, High-achieving group) showed that this group managed to reach a consensus, despite initial confusion [CO] (Fig. 3, top). The students brainstormed ideas to determine the location of a crash site without using devices. Student A suggested that the crash site was in North America [NK]. Student B confirmed and provided more specific information, stating that they were in the US [CK]. Student A's uncertainty and confusion caused a pause [CO], but the discussion resumed with confirming the ideas [CK]. Student C suggested seeking assistance from the instructors, while D explored alternative solutions [NK]. It is noteworthy that despite the extended pause, the group persisted with the idea put forth by the group, demonstrating collective focus.

Vignette 4 (Task-4, Low-achieving group) shows a group struggling with joint attention, leading to inadequate uptake and integration of ideas (Fig. 3, bottom). Student A discovered the declination angle of Polaris [NK] but did not foster overlapping scenes to share findings. Student B was confused about what A had just shared [CO]. Student A used gestures to explain the meaning of declination [NK], but B remained confused [CO]. Meanwhile, Student C, who was also exploring the task individually while wearing an AR headset, joined the conversation but suggested following what they had explored instead of what A had suggested. This divergent focus and poor information sharing left Student B perplexed.

4.3.2. ONA analyses

The vignettes provided a rich, narrative-driven examination of the differences in multimodal interactions between low- and high-achieving groups. Building on these qualitative insights, we recognized the need for a more granular analysis to uncover feature connections. Thus, we employed ONA to investigate the co-occurrence and directionality between students' multimodal interactions (i.e., five JA states and four conceptual discussion codes) across different tasks. ONA network graphs were first created for each task and performance group, providing detailed insights into node connections and network structure (see Fig. 4). Additionally, we created subtracted ONA networks to highlight performance group differences. The size of the colored circles within nodes represents feature duration and frequency. Directed connections are depicted as edges, with chevrons indicating connection direction to a subsequent JA state or discussion code.

The Task-1 graphs indicate that the low and high-achieving groups have similar network structures. This is shown by the comparable connection patterns in both individual networks (left and right) and the light edge weights and overlapping projected mean points in the subtracted network (center). For example, most of the points are located in the lower right of the network except few outliers and the edges are similar between the two networks. Both performance categories display a strong connection between *JA II*

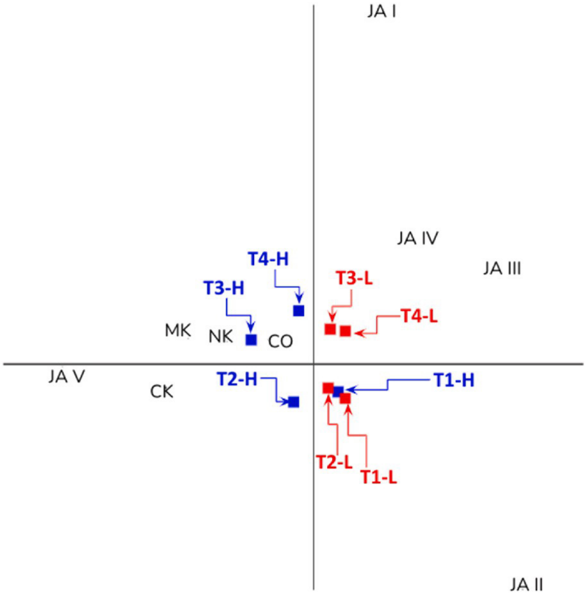


Fig. 5. ONA projected mean squares. Each square summarizes a representation of the network within the specific performance group (high: blue; low: red). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

and *confirm knowledge* and negligible engagement with *JA V*. These patterns are not surprising as students were expected to explore the simulation environment individually in the early problem-solving stage.

During Task-2, both networks maintain a strong connection with *JA II* and *confirm knowledge*. The subtracted network reveals that the high-achieving groups engaged in *confirming knowledge* with *JA IV* and *JA V*. Recall that these JA states indicate two devices were looking at different or overlapped areas in Horizon, representing more proactive coordination behaviors. Specifically, the connection directionality implies that high-achieving groups confirm their understanding and then coordinate their device screens accordingly for visual reference. In contrast, the network of low-achieving groups showed minimal difference from Task-1, consistently maintaining lower levels of visual synchronization and coordination.

During Task-3, the high-achieving groups demonstrated a stronger connection between *JA V* and *confirm knowledge*, a pattern unobserved in low-achieving groups. Additionally, the size of the colored circle in *JA V* is proportionally large in the high-achieving network, indicating their sustained engagement in the JA state. These findings imply the ability of the high-achieving groups to build a shared problem space through verbal communication (physical space) and digital interactions (virtual space). The ONA graphs from the last two tasks reveal a strong connection between *JA I* and *confirm knowledge* for both high- and low-achieving groups, but with opposite directionality. A sequential trend from *JA I* to *confirm knowledge* suggests that the high-achieving groups engaged in conceptual discussions without technology, subsequently leading to knowledge confirmation. This pattern aligned with our vignette analyses. The last task required the most complex mathematical computations, which in turn stimulated more verbal exchanges.

Lastly, Fig. 5 shows the summary ONA network graph across all tasks. Given the MR technique (see 3.5), we focused on interpreting the first dimension (x-axis) as it accounted for a significant amount of explained variance. This dimension clearly differentiates connections related to two modalities: JA states and conceptual discussion. Nodes positioned on the left side, particularly *JA V*, exhibit the highest level of synchrony in joint attention, while *JA II* on the right side represents minimal synchrony within the simulation. Notably, all conceptual discussion codes are grouped together on the left, near *JA V*, indicating their inherent connection. We interpret the x-axis as representing the degree of behavioral synchrony in attention coordination and its relation to conceptual discussion within the network.

The eight squares represent the mean points of two performance categories and quantitatively capture the average connection strength among all nodes for each task. Interestingly, the low and high-achieving groups show similar associations between JA states and conceptual interactions for Task-1 (T1-H and T1-L). However, different connection patterns emerge for the subsequent, more complex tasks. The squares on the left side of the space, representing high-achieving groups, are closer to all conceptual discussion nodes and *JA V*, while squares on the right side are closer to *JA III* and *JA IV*. This indicates the high-achieving groups engaged more with *JA V* and conceptual discussion during the later tasks.

5. Discussion

This paper aimed to advance methods to analyze joint attention as a key indicator to unpack complex collaborative learning processes in an immersive learning environment. The open-ended nature of immersive environments poses challenges in tracing individual actions and collective activity. By examining joint attention as a multiple-level indicator of collaboration, our study shed light on how students collaborate with peers within and beyond the simulation. Our approach goes beyond traditional analysis by incorporating multimodal data, illustrating the intricate relationship between knowledge construction and various forms of student multimodal engagement, spanning both digital and cognitive spaces. By adopting a process-oriented perspective (e.g., Janssen et al., 2010), we introduced an innovative multimodal temporal approach to analyzing students' interactions using ONA. This approach offers a more dynamic and comprehensive view of evolving multimodal engagement patterns and their association with learning performances.

Our findings from RQ1 highlight the limitation of using aggregated measures of JA states to distinguish meaningful differences in attention coordination and their relationship with collaborative learning performance. In RQ2, we supplemented joint attention with two additional modalities–*conceptual turn-taking* and *gesture*, recognized as essential facets of collaboration (Jermann & Nüssli, 2012). The inclusion of *conceptual turn-taking* improved the understanding of group dynamics, however, adding *gestures* did not provide significant insights beyond that. While these investigations yielded valuable insights into incorporating multimodal features for a more nuanced understanding of collaboration, they did not fully address the complexity of collaboration, particularly the interconnected aspects that evolve over time. In RQ3, we, therefore, delved deeper into the process of group knowledge construction by analyzing the selected vignettes, and further by utilizing ONA, an approach emphasizing temporal dimensions of co-occurrence and sequentiality. Multiple network graphs visualized the interaction patterns distinguishing low and high-achieving groups. The summary and task-level graphs captured variations in engagement with JA states and conceptual discussion types, revealing how these differences change over time.

5.1. Theoretical implications

Incorporating additional modalities, including audio and gesture data, enabled a more nuanced understanding of the role of joint attention in group sense-making (Blikstein & Worsley, 2016; Schneider et al., 2021). This study highlights the need for a multidimensional lens to understand various intertwined aspects that evolve over time (Dillenbourg et al., 1996) and adds a new layer of complexity to our existing knowledge of joint attention. While the binary representation of joint attention in previous studies successfully revealed its association with collaboration, they lacked the ability to explain how and when joint attention contributes to collaboration. Our approach, by depicting multimodal interaction patterns with ONA networks (Fig. 4), we discovered that joint attention facilitates knowledge construction in various ways. For example, it offers a visual reference through screen coordination. This visual synchronization helps reduce misunderstanding and foster clarity, thereby supporting knowledge confirmation. It also cultivates sustained engagement in shared attention within virtual space, fostering higher-quality discussions and more efficient problem-solving in cognitive space.

Our multimodal temporal approach sheds light on distinct interaction patterns across varying performance categories and tasks, highlighting the importance of attention regulation and coordination. Overall, high-achieving groups exhibited higher degrees of joint attention, underscoring its importance in productive collaboration (Richardson et al., 2017). Both low- and high-achieving groups exhibited comparable interaction patterns at their early problem-solving stage. This similarity began to diverge as high-achieving groups developed stronger connections with shared attention, while low-achieving groups maintained lower joint attention levels mostly across all tasks. These observations suggest that joint attention does not need to be sustained all the time; instead, it should be gained in critical moments (Barron, 2003). This insight enriches our understanding of joint attention and supports the notion of timely joint attention. It also pinpoints the need for additional support to regulate group attention and establish visual coordination during moments when shared attention becomes particularly important (Teasley & Roschelle, 1993).

Another distinguishing factor between low- and high-achieving groups is associated with the role of individual exploration, which is reflected in the lower JA state features observed during the early stages. This indicates divergence in group attention and echoes the notion of idea divergence (Kapur et al., 2008) and divergent inquiry (Tissenbaum et al., 2017). Groups could potentially gain advantages from such a divergence during the initial stages of problem-solving, as it may facilitate the exploration and acquisition of new information (Tissenbaum, 2020). Our findings revealed that high-achieving groups exhibited a pattern of initial divergence followed by attention convergence, characterized by higher degrees of joint attention and conceptual discussions, during subsequent tasks. This transition may serve as a key mechanism facilitating the exchange of information and construction of collective knowledge. Future research is needed to determine whether such transitions are predominantly observed within open-ended learning environments that foster exploration, or if they can also manifest in other types of collaborative learning environments.

5.2. *Methodological implications*

The analytical decisions and methodological design adopted in our study carry several implications. First, it provides empirical evidence underscoring the benefits of combining features obtained from multiple data sources, contributing to a more holistic understanding of collaborative learning outcomes (Olsen et al., 2020; Worsley & Blikstein, 2018). Our investigation into RQ2 suggests incorporating turn-taking, an index of verbal engagement, better contextualizes the association between joint attention and knowledge co-construction. This additional audio modality offers a supplementary viewpoint, capturing group interactions beyond the digital interfaces, thus providing a more holistic view of the collaboration process across both digital and cognitive spaces.

Second, this study highlights the advantages of analyzing data at multiple temporal granularities (all-tasks, task, and episode levels), with each providing unique insights. The correlation analysis (macro-level) revealed a positive relationship among screen coordination behavior, verbal interaction, and collective knowledge construction. However, this approach fell short of capturing the evolving nature of the knowledge-building process or fine-grained patterns of interaction (Reimann, 2009). To address this, we employed ONA (meso-level) and vignette analysis (micro-level) to unpack sequential and contemporaneous relationships. The summary ONA network served as a critical tool in revealing distinct interaction patterns among varying performance categories and tasks. It showed advanced coordination behaviors in high-achieving groups, especially during later, more complex tasks. Individual ONA networks offered a more nuanced understanding of directional connection, illustrating how JA states catalyze or are guided by knowledge construction codes. These approaches offer different levels of understanding: macro-level provides an overview, meso-level tracks temporal trends, and micro-level uncovers intricate patterns. Future research should adopt multimodal analysis at multiple granular levels to capture the richness and complexity of collaborative learning comprehensively.

Finally, we showcase the importance of *closing the interpretative loop* in research design, a term emphasizing transparency and traceability by grounding complex models in original narratives (Shaffer, 2017). We began by deriving concrete insights from the narrative-driven vignettes, setting a foundation that informed and inspired our subsequent ONA analyses. The ONA allowed us to validate and generalize our qualitative findings, ensuring that the granular behaviors observed in the vignettes were consistently reflected at a macro-level analysis. This method resulted in a deeper comprehension of the underlying mechanisms of joint attention and further collaborative learning.

5.3. *Instructional implications*

Our findings suggest several implications from three dimensions essential for the instruction and design of collaborative learning environments: task, tool, and instructors (Mercier et al., 2023). Firstly, our observation of high-achieving groups exhibiting early divergence followed by convergence suggests that tasks should be crafted to promote individual exploration and diverse perspectives in the early stage of collaboration. This approach enables students to cultivate their own ideas and interpretations. Moreover, tools can offer resources and prompts in guiding the transition from individual exploration to collective discussion (Kapur et al., 2008; Tissenbaum, 2020). Instructors should actively foster group communication that integrates diverse perspectives, as this can yield improved coordination, error identification, and future planning. This also underscores the orchestration goals for instructors, which emphasize the significance of nurturing effective communication, shared understanding, and attention management skills (Fiore et al., 2017; Ja¨rvel¨a & Hadwin, 2013; Van de Pol et al., 2010). Second, the sustained lower joint attention levels observed in the low-achieving groups point to challenges in cognitive processing and attention coordination due to competing demands on focus

(Borge et al., 2022). It is critical that tasks be carefully designed to reduce demands and distractions from task-irrelevant activities. Lastly, our research highlights the importance of promoting collective attention and shared understanding timely (Barron, 2003). Group awareness tools that track students' multimodal engagement, including focus areas and participant contributions, can offer timely feedback at critical moments (Buder et al., 2022; J¨arvela¨ & Hadwin, 2013). These tools not only help the group stay on track but also empower instructors to enhance collective experiences.

5.4. Limitations

We acknowledge several limitations. First, the small sample size, containing only 16 groups from one university, may limit the generalizability of our findings. Second, our selection of multimodal features was guided by our research questions and specific context, potentially leaving out other important aspects of collaborative learning. Additional behavioral traces and affective states could reveal further insights into collaborative learning in immersive environments. Furthermore, our approach to assessing inter-rater reliability based on percent agreement may be susceptible to Type 1 error (Eagan et al., 2020). This suggests the importance of employing more robust methods in evaluating inter-rater reliability. Lastly, the ONA required specifying modeling parameters like *window size*. While we used statistical tests to determine the one with the best model fit and variance explained, determining the optimal parameter remains an ongoing research.

6. Conclusions

This research makes important theoretical and methodological contributions to understanding complex collaboration dynamics in an immersive learning environment. Our approach introduces multiple-level joint attention indicators to measure varying degrees of attention coordination across devices, alongside a multimodal temporal approach to enhance the current understanding of collaborative learning. Our research uncovers limitations of aggregated measures to distinguish differences in attention coordination behaviors and the importance of complimenting joint attention with other multimodal traces, allowing us to capture group interactions beyond digital interfaces. This broader perspective provides a more holistic view of the collaboration process across digital and cognitive spaces. Theoretical implications, obtained through statistical, qualitative, and ordered network analyses, include the notion of timely joint attention (Barron, 2003) and empirical evidence for the potential benefits of initial attention divergence for knowledge convergence in later stages (Kapur et al., 2008). Our multi-granular methodological approach leveraging macro-level correlation, meso-level network analysis, and micro-level qualitative vignettes provides a comprehensive view to showcase the importance of grounding computational modeling in qualitative narratives and advance the understanding of joint attention's role across tasks. Such empirical evidence has practical implications for instructors and designers, especially in critical moments, requiring shared attention during collaborative problem-solving. Recognizing the lack of frameworks characterizing joint attention, this study underscores the critical need to identify how and when joint attention promotes collaboration. By considering temporality and multimodality, this study significantly contributes to a growing demand for a deeper understanding of the complex dynamics of collaborative learning and how to support productive student interactions.

CRediT authorship contribution statement

Jina Kang: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Yiqiu Zhou:** Formal analysis, Methodology, Visualization, Writing - original draft, Writing - review & editing. **Robin Jephthah Rajarathinam:** Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Yuanru Tan:** Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **David Williamson Shaffer:** Supervision, Conceptualization.

Declaration of competing interest

We declare no competing interest.

Data availability

Data will be made available on request.

Acknowledgements

We would like to thank Nathan Kimball and his team at Concord Consortium for developing the software that made this research possible. We would also like to thank the entire CEASAR staff for their invaluable support in conducting this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compedu.2024.105002.

References

Abitino, A., Pugh, S. L., Peacock, C. E., & D'Mello, S. K. (2022). Eye to eye: Gaze patterns predict remote collaborative problem solving behaviors in triads. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial intelligence in education* (Vol. 13355, pp. 378–389). Springer International Publishing. https://doi.org/10.1007/978-3-031-11644-5_31.

Akçayır, M., & Akçayır, G. (2017). Advantages and challenges associated with augmented reality for education: A systematic review of the literature. *Educational Research Review, 20*, 1–11.

Albanese, A., Danhoni Neves, M. C., & Vicentini, M. (1997). Models in science and in education: A critical review of research on students' ideas about the Earth and its place in the universe. *Science & Education, 6*(6), 573–590.

Bakeman, R., & Adamson, L. B. (1984). Coordinating attention to people and objects in mother-infant and peer-infant interaction. *Child Development, 55*, 1278–1289.

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

Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics, 3*(2), 220–238.

Borge, M., Aldemir, T., & Xia, Y. (2022). How teams learn to regulate collaborative processes with technological support. *Educational Technology Research & Development, 70*(3), 661–690.

Bovo, R., Giunchi, D., Alebri, M., Steed, A., Costanza, E., & Heinis, T. (2022). Cone of vision as a behavioural cue for VR collaboration. In *Proceedings of the ACM on human-computer interaction* (Vol. 6, pp. 1–27). https://doi.org/10.1145/3555615. CSCW2.

Bowman, D., Swiecki, Z., Cai, Z., Wang, Y., Eagan, B., Linderoth, J., & Shaffer, D. W. (2021). The mathematical foundations of epistemic network analysis. In *Advances in quantitative ethnography: Second international conference, ICQE 2020* (Vol. 2, pp. 91–105). Malibu, CA, USA: Springer International Publishing. *February 1-3, 2021, Proceedings*.

Brenner, C., DesPortes, K., Hendrix, J. O., & Holford, M. (2021). GeoForge: Investigating integrated virtual reality and personalized websites for collaboration in middle school science. *Information and Learning Sciences, 122*(7/8), 546–564.

Buder, J., Bodemer, D., & Ogata, H. (2022). Group awareness. In U. Cress, C. Ros´e, A. F. Wise, & J. Oshima (Eds.), *International handbook of computer-supported collaborative learning* (pp. 295–313). Cham: Springer.

Chung, C. Y., Awad, N., & Hsiao, I. H. (2021). Collaborative programming problem-solving in augmented reality: Multimodal analysis of effectiveness and group collaboration. *Australasian Journal of Educational Technology, 37*(5), 17–31.

Cukurova, M., Giannakos, M., & Martinez-Maldonado, R. (2020). The promise and challenges of multimodal learning analytics. *British Journal of Educational Technology, 51*(5), 1441–1449. https://doi.org/10.1111/BJET.13015

Di Mitri, D., Schneider, J., Klemke, R., Specht, M., & Drachsler, H. (2019). Read between the lines: An annotation tool for multimodal data for learning. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 51–60).

Diederich, M., Kang, J., Kim, T., & Lindgren, R.. Developing an in-application shared view metric to capture collaborative learning in a multi-platform astronomy simulation. https://doi.org/10.1145/3448139.3448156.

Dillenbourg, P. (1999). Introduction: What do you mean by "collaborative learning"? In P. Dillenbourg (Ed.), *Collaborative learning: Cognitive and computational approaches* (pp. 1–19). Amsterdam: Pergamon.

Dillenbourg, P., Baker, M., Blaye, A., & O'Malley, C. (1996). The evolution of research on collaborative learning. In E. Spada, & P. Reiman (Eds.), *Learning in Humans and Machine: Towards an interdisciplinary learning science* (pp. 189–211). Oxford: Elsevier.

Dunleavy, M., & Dede, C. (2014). Augmented reality teaching and learning. *Handbook of research on educational communications and technology*, 735–745.

Eagan, B., Brohinsky, J., Wang, J., & Shaffer, D. W. (2020). Testing the reliability of inter-rater reliability. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 454–461).

Fan, Y., Tan, Y., Rakovi´c, M., Wang, Y., Cai, Z., Shaffer, D. W., & Gaˇsevi´c, D. (2023). Dissecting learning tactics in MOOC using ordered network analysis. Journal of Computer Assisted Learning, 39(1), 154-166.

Fiore, S. M., Graesser, A., Greiff, S., Griffin, P., Gong, B., Kyllonen, P., … von Davier, A. (2017). *Collaborative problem solving: Considerations for the national assessment of educational progress*. Alexandria, VA: National Center for Education Statistics.

Giannakos, M., Cukurova, M., & Papavlasopoulou, S. (2022). Sensor-based analytics in education: Lessons learned from research in multimodal learning analytics. In *Multimodal learning analytics handbook* (pp. 329–358). Cham: Springer International Publishing.

Goldhammer, F., Naumann, J., Stelter, A., To´th, K., Ro¨lke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608.

Hahn, L., & Klein, P. (2022). Eye tracking in physics education research: A systematic literature review. *Physical Review Physics Education Research, 18*(1), Article 013102.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64–74.

Hmelo-Silver, C. E., & Barrows, H. S. (2008). Facilitating collaborative knowledge building. *Cognition and Instruction, 26*(1), 48–94.

Iba´n˜ez, M. B., Di Serio, A´., Villara´n, D., & Kloos, C. D. (2014). Experimenting with electromagnetism using augmented reality: Impact on flow student experience and educational effectiveness. *Computers & Education, 71*, 1–13.

Janssen, J., Kirschner, F., Erkens, G., Kirschner, P. A., & Paas, F. (2010). Making the black box of collaborative learning transparent: Combining process-oriented and cognitive load approaches. *Educational Psychology Review, 22*, 139–154.

Ja¨rvel¨a, S., & Hadwin, A. F. (2013). New frontiers: Regulating learning in CSCL. *Educational Psychologist, 48*(1), 25–39.

Jermann, P., Mullins, D., Nüssli, M. A., & Dillenbourg, P. (2011). *Collaborative gaze footprints: Correlates of interaction quality*.

Jermann, P., & Nuessli, M. A. (2011). Unraveling cross-recurrence: Coupling across timescales. In *Proceedings of international workshop on dual eye tracking in CSCW (DUET 2011)*. Aarhus.

Jermann, P., & Nüssli, M. A. (2012). Effects of sharing text selections on gaze cross-recurrence and interaction quality in a pair programming task. In *Proceedings of the ACM 2012 conference on computer supported cooperative work* (pp. 1125–1134).

Johri, A., Williams, C., & Pembridge, J. (2013). Creative collaboration: A case study of the role of computers in supporting representational and relational interaction in student engineering design teams. *International Journal of Engineering Education, 29*(1), 33–44.

Kapur, M., Voiklis, J., & Kinzer, C. K. (2008). Sensitivities to early exchange in synchronous computer-supported collaborative learning (CSCL) groups. *Computers & Education, 51*(1), 54–66.

L¨amsa¨, J., Kotkajuuri, J., Lehtinen, A., Koskinen, P., M¨antyl¨a, T., Kilpela¨inen, J., & H¨am¨ala¨inen, R. (2022). The focus and timing of gaze matters: Investigating collaborative knowledge construction in a simulation-based environment by combined video and eye tracking. *Frontiers in Education, 7*, Article 942224. https://doi.org/10.3389/feduc.2022.942224

Liu, C. C., Hsieh, I. C., Wen, C. T., Chang, M. H., Chiang, S. H. F., Tsai, M. J., … Hwang, F. K. (2021). The affordances and limitations of collaborative science simulations: The analysis from multiple evidences. *Computers & Education, 160*, Article 104029.

Ludvigsen, S., Cress, U., Ros´e, C. P., Law, N., & Stahl, G. (2018). Developing understanding beyond the given knowledge and new methodologies for analyses in CSCL. *International Journal of Computer-Supported Collaborative Learning, 13*, 359–364. https://doi.org/10.1007/s11412-018-9291-0

Manathunga, K., & Hern´andez-Leo, D. (2015). Has research on collaborative learning technologies addressed massiveness? A literature review. *Journal of Educational Technology & Society, 18*(4), 357–370.

Marquart, C., Tan, Y., Cai, Z., & Shaffer, D. (2022). ona: Ordered network analysis. R package version 0.1.1.1684949787 https://cran.qe-libs.org/ona/index.html.

Martinez-Maldonado, R., Kay, J., Buckingham Shum, S., & Yacef, K. (2019). Collocated collaboration analytics: Principles and dilemmas for mining multimodal interaction data. *Human-Computer Interaction, 34*(1), 1–50. https://doi.org/10.1080/07370024.2017.1338956

Melzner, N., Greisel, M., Dresel, M., & Kollar, I. (2019). Using process mining (PM) and epistemic network analysis (ENA) for comparing processes of collaborative problem regulation. In B. Eagan, M. Misfeldt, & A. Siebert-Evenstone (Eds.), *Advances in quantitative ethnography, communications in computer and information science* (pp. 154–164). Springer International Publishing.

Mercier, E., Vourloumi, G., & Higgins, S. (2017). Student interactions and the development of ideas in multi-touch and paper-based collaborative mathematical problem solving. *British Journal of Educational Technology, 48*(1), 162–175, 10/f9rd4h.

Mercier, E., Goldstein, M. H., Baligar, P., & Rajarathinam, R. J. (2023). Collaborative Learning in Engineering Education. In International Handbook of Engineering Education Research (pp. 402-432). Routledge.

Moore, C., & Dunham, P. J. (Eds.). (1995). *Joint attention: Its origins and role in development*. Lawrence Erlbaum Associates, Inc.

Olsen, J. K., Sharma, K., Rummel, N., & Aleven, V. (2020). Temporal analysis of multimodal data to predict collaborative learning outcomes. *British Journal of Educational Technology, 51*(5), 1527–1547.

O'Madagain, C., & Tomasello, M. (2021). Joint attention to mental content and the social origin of reasoning. *Synthese, 198*(5), 4057–4078. https://doi.org/10.1007/s11229-019-02327-1

Palincsar, A. S. (1998). Social constructivist perspectives on teaching and learning. *Annual Review of Psychology, 49*(1), 345–375. https://doi.org/10.1146/annurev.psych.49.1.345

Planey, J., Rajarathinam, R. J., Mercier, E., & Lindgren, R. (2023). Gesture-mediated collaboration with augmented reality headsets in a problem-based astronomy task. *International Journal of Computer-Supported Collaborative Learning, 18*(2), 259–289.

Reimann, P. (2009). Time is precious: Variable-and event-centered approaches to process analysis in CSCL research. *International Journal of Computer-Supported Collaborative Learning, 4*, 239–257.

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science, 29*(6), 1045–1060.

Richardson, D. C., Dale, R., & Kirkham, N. Z. (2007). The art of conversation is coordination. *Psychological Science, 18*(5), 407–413.

Ruis, A., Siebert-Evenstone, A., Pozen, R., Eagan, B., & Shaffer, D. (2019). Finding common ground: A method for measuring recent temporal context in analyses of complex, collaborative thinking. In K. Lund, G. P. Niccolai, E. Lavou´e, C. Hmelo-Silver, G. Gweon, & M. Baker (Eds.), *A wide lens: Combining embodied, enactive, extended, and embedded Learning in collaborative settings, 13th international conference on computer supported collaborative learning (CSCL) 2019, ume 1* pp. 136–143). Lyon, France: International Society of the Learning Sciences.

Saint, J., Fan, Y., Gaˇsevi´c, D., & Pardo, A. (2022). Temporally-focused analytics of self-regulated learning: A systematic review of literature. *Computers and Education: Artificial Intelligence, 3*, Article 100060.

Schneider, B., & Bryant, T. (2022). Using mobile dual eye-tracking to capture cycles of collaboration and cooperation in Co-located dyads. *Cognition and Instruction*, 1–30. https://doi.org/10.1080/07370008.2022.2157418

Schneider, B., & Pea, R. (2013). Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *International Journal of Computer-Supported Collaborative Learning, 8*(4), 375–397.

Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., & Pea, R. (2016). Using mobile eye-trackers to unpack the perceptual benefits of a tangible user interface for collaborative learning. *ACM Transactions on Computer-Human Interaction, 23*(6), 1–23.

Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., & Pea, R. (2018). Leveraging mobile eye-trackers to capture joint visual attention in co-located collaborative learning groups. *International Journal of Computer-Supported Collaborative Learning, 13*, 241–261.

Schneider, B., Sung, G., Chng, E., & Yang, S. (2021). How can high-frequency sensors capture collaboration? A review of the empirical links between multimodal metrics and collaborative constructs. *Sensors, 21*(24), 8185.

Scollon, R., & Scollon, S. (2009). Multimodality and language: A retrospective and prospective view. In C. Jewitt (Ed.), *The routledge handbook of multimodal analysis* (pp. 170–180). London: Routledge.

Shaffer, D. W. (2017). *Quantitative ethnography*. Madison, WI: Cathcart Press.

Sharma, K., & Giannakos, M. (2020). Multimodal data capabilities for learning: What can multimodal data tell us about learning? *British Journal of Educational Technology, 51*(5), 1450–1484.

Sharma, K., Jermann, P., Dillenbourg, P., Prieto, L. P., D'Angelo, S., Gergle, D., Schneider, B., Rau, M., Pardos, Z., & Rummel, N. (2017). CSCL and eye-tracking: Experiences, opportunities and challenges. In B. K. Smith, M. Borge, E. Mercier, & K. Y. Lim (Eds.), *Making a difference: Prioritizing equity and access in CSCL, 12th international conference on computer supported collaborative learning (CSCL) 2017* (Vol. 2). International Society of the Learning Sciences.

Shehab, S., & Mercier, E. (2020). Exploring the relationship between the types of interactions and progress on a task during collaborative problem solving. In I. S. Horn, & M. Gresalfi (Eds.), *The interdisciplinarity of the learning sciences, 14th international conference of the learning sciences* (Vol. 3, pp. 1285–1292). International Society of the Learning Sciences (ISLS). https://repository.isls.org//handle/1/6326.

Siposova, B., & Carpenter, M. (2019). A new look at joint attention and common knowledge. *Cognition, 189*, 260–274.

Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge (acting with technology)*. The MIT Press.

Swiecki, Z., Ruis, A. R., Farrell, C., & Shaffer, D. W. (2020). Assessing individual contributions to collaborative problem solving: A network analysis approach. *Computers in Human Behavior, 104*, Article 105876.

Tan, Y., Ruis, A. R., Marquart, C., Cai, Z., Knowles, M. A., & Shaffer, D. W. (2022, October). Ordered network analysis. In International Conference on Quantitative Ethnography (pp. 101-116). Cham: Springer Nature Switzerland.

Teasley, S. D., & Roschelle, J. (1993). Constructing a joint problem space: The computer as a tool for sharing knowledge. *Computers as cognitive tools*, 229–258.

Tissenbaum, M. (2020). I see what you did there! Divergent collaboration and learner transitions from unproductive to productive states in open-ended inquiry. *Computers & Education, 145*, Article 103739.

Tissenbaum, M., Berland, M., & Lyons, L. (2017). DCLM framework: Understanding collaboration in open-ended tabletop learning environments. *International Journal of Computer-Supported Collaborative Learning, 12*, 35–64.

Van de Pol, J., Mercer, N., & Volman, M. (2019). Scaffolding student understanding in small-group work: Students' uptake of teacher support in subsequent small-group interaction. *The Journal of the Learning Sciences, 28*(2), 206–239. https://doi.org/10.1080/10508406.2018.1522258

Villanueva, A., Zhu, Z., Liu, Z., Peppler, K., Redick, T., & Ramani, K. (2020). Meta-AR-app: An authoring platform for collaborative augmented reality in STEM classrooms. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–14).

Vrzakova, H., Amon, M. J., Stewart, A., Duran, N. D., & D'Mello, S. K. (2020). 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).

Webb, M., Tracey, M., Harwin, W., Tokatli, O., Hwang, F., Johnson, R., … Jones, C. (2022). Haptic-enabled collaborative learning in virtual reality for schools. *Education and Information Technologies, 27*(1), 937–960.

Wisiecka, K., Konishi, Y., Krejtz, K., Zolfaghari, M., Kopainsky, B., Krejtz, I., Koike, H., & Fjeld, M. (2023). Supporting complex decision-making. Evidence from an eye tracking study on in-person and remote collaboration. *ACM Transactions on Computer-Human Interaction*. https://doi.org/10.1145/3581787

Worsley, M., & Blikstein, P. (2018). A multimodal analysis of making. *International Journal of Artificial Intelligence in Education, 28*, 385–419.

Zhou, R., & Kang, J. (2022). Characterizing joint attention dynamics during collaborative problem-solving in an immersive astronomy simulation. In *Proceedings of the 15th International Conference on Educational Data Mining* (pp. 406–413). https://doi.org/10.5281/zenodo.6852988

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods, 12*(4), 399.