JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 10(3), 87–101. https://doi.org/10.18608/jla.2023.7989

Enriching Multimodal Data: A Temporal Approach to Contextualize Joint Attention in Collaborative Problem-Solving

Yiqiu Zhou[1] and Jina Kang[2]

Abstract

Collaboration is a complex, multidimensional process; however, details of how multimodal features intersect and mediate group interactions have not been fully unpacked. Characterizing and analyzing the temporal patterns based on multimodal features is a challenging yet important work to advance our understanding of computer-supported collaborative learning (CSCL). This paper highlights the affordances, as well as the limitations, of different temporal approaches in terms of analyzing multimodal data. To tackle the remaining challenges, we present an empirical example of multimodal temporal analysis that leverages multi-level vector autoregression (mlVAR) to identify temporal patterns of the collaborative problem-solving (CPS) process in an immersive astronomy simulation. We extend previous research on joint attention with a particular focus on the added value from a multimodal, temporal account of the CPS process. We incorporate verbal discussion to contextualize joint attention, examine the sequential and contemporaneous associations between them, and identify significant differences in temporal patterns between low- and high-achieving groups. Our paper does the following: 1) creates interpretable multimodal group interaction patterns, 2) advances understanding of CPS through examination of verbal and non-verbal interactions, and 3) demonstrates the added value of a complete account of temporality including both duration and sequential order.

Notes for Practice

- Incorporating temporal and multimodal aspects into the modelling of collaboration presents conceptual and practical challenges. This paper demonstrates the value of analyzing temporal structure and interdependence among multimodal features.
- Leveraging a multimodal temporal approach, this work reveals differences in temporal dynamics of verbal and non-verbal interactions between low and high-achieving groups.
- Key findings characterize productive collaboration through sustained joint attention, individual exploration triggering discussion, and concurrent digital-conceptual engagement. These insights can inform the design of collaborative learning environments that facilitate individual reflection for knowledge discovery and guide joint coordination at critical moments for knowledge exchange.

Keywords

Collaborative problem-solving, temporality, temporal analysis, multimodal learning analytics, collaboration analytics, CSCL

Submitted: 07/03/2023 — **Accepted:** 03/08/2023 — **Published:** 04/11/2023

Corresponding author [1]Email: yiqiuz3@illinois.edu Address: Department of Curriculum & Instruction, College of Education, University of Illinois Urbana-Champaign, 1310 S. Sixth St., Champaign, IL 61820-6925. ORCID iD: https://orcid.org/0009-0000-9004-9307
[2]Email: jinakang@illinois.edu Address: Department of Curriculum & Instruction, College of Education, University of Illinois Urbana-Champaign, 1310 S. Sixth St., Champaign, IL 61820-6925. ORCID iD: https://orcid.org/0000-0002-7043-5427

1. Introduction

Computer-supported collaborative learning (CSCL) environments have the potential to facilitate joint activity and support collaborative problem-solving (CPS) by creating opportunities for students to construct shared knowledge (Dillenbourg et al., 2009). Advanced technologies, such as virtual and augmented reality, open up new possibilities for collaborative learning through their unique affordances to add virtual objects to the real world and support different types of interactions and communication cues (Phon et al., 2014). Yet, these immersive environments also place emergent demands to model and

analyze the collaboration process due to the complexity of these multimodal interactions, such as speech, gesture, and touch (Nizam et al., 2018). New research interests began to focus on multimodal learning analytics (MMLA) as a potential avenue to monitor, analyze, and model interactions in the collaborative learning process (Cukurova et al., 2020; Praharaj et al., 2021). Although research on MMLA is of significant value, it is confronted with challenges. When multiple data sources were fused to model or predict collaboration, most previous studies employed a "naïve" fusion approach (Chango et al., 2022; Worsley, 2014) — features extracted from various data sources were aggregated without a hypothesis about how they might interact. This data fusion approach could be problematic for two reasons: 1) it ignores the temporal variation of each feature and assumes temporal homogeneity (Kapur et al., 2008; Worsley, 2014), which is usually inconsistent with the evolving, dynamic nature of collaboration, and 2) it leads to difficulty in interpreting these features in terms of their relations to and effects on learning (Noroozi et al., 2019).

This lack of temporal consideration limits the application and expansion of previous empirical studies since it is unclear how various collaboration measures interreact to mediate collaboration processes. For instance, a recent study by Emerson et al. (2020) found that predictive models based on two modalities outperformed those of three modalities, suggesting a need to further unpack how multiple features contribute to modelling collaboration. As such, we argue that both the empirical literature and practical research in MMLA could benefit from an additional account of temporality. This process-oriented view could help researchers make explicit analytical decisions, such as whether to focus on duration, sequence, and change over time, as well as the integration between temporality and the targeted collaboration constructs (Chen et al., 2018). It also opens up a new possibility to re-examine the temporal structure and interdependence relationship of multimodal features to interpret the collaboration process, potentially leading to more generalized conclusions and new research questions and insights.

Thus, this paper highlights the importance of a process-oriented perspective to account for both temporality and multimodality. Greater attention should be given to understanding how features extracted from multiple data sources intersect and mediate interaction in the collaboration process. The purpose of this work is to present an empirical example that enables a temporal analysis of multimodal features extracted from different data sources. Instead of modelling the collaboration process by aggregating all its features, we aim to unpack its black box by examining the dependent temporal structure and contemporaneous correlations between different features. This approach allows us to understand how verbal and non-verbal features intertwine and ultimately lead to effective collaboration. The remaining sections are organized as follows. We first review temporal analysis in CSCL, along with its affordances and challenges. Computational approaches that support the temporal analysis of multimodal data are discussed, with a focus on data representations and assumptions regarding temporal relationships. Next, we introduce multi-level vector autoregression (mlVAR) as a time series analysis to capture the interplays between verbal and non-verbal interactions in an immersive collaborative learning environment. Lastly, we end by discussing the findings and implications for conducting temporal analysis on multimodal data.

2. Relevant Work

2.1. Temporal Analysis in CSCL

Temporal analysis has been broadly defined as approaches focusing on temporal characteristics of events or interrelations between these events (Lämsä et al., 2021). There has been increasing awareness that temporality is a key characteristic of CSCL as learning is a process unfolding over time (Knight et al., 2017; Reimann, 2009). Researchers began to explore temporal patterns in group interactions and how these patterns influence collaboration quality and/or learning performance. Despite the progress made so far, the field still faces the following challenges when employing temporal analysis: 1) incomplete conceptual understanding of temporality, 2) limited insights in unimodal data, 3) contextualization issues, and 4) difficulty determining an appropriate analytical unit.

The first issue in practice is to incorporate a comprehensive conceptual understanding of temporality into analytic approaches. The concept of temporality contains both the passage of time and the sequential order of events (Chen et al., 2018; Molenaar & Wise, 2022). The passage of time emphasizes the continuous flow while the sequential order focuses on the relative arrangement of multiple events. However, in many cases, only one aspect of temporal features is highlighted, assuming that different events and behaviours take place in the same order or have the same flow (Chen et al., 2017; Knight et al., 2017). Inferential statistics and coding-and-accounting approaches, for example, ignore the sequential order and assume temporal homogeneity, which has been proven rarely valid (Kapur et al., 2008) and thus inadequate to capture the dynamics of group interactions (Lämsä et al., 2021). Sequential analysis acknowledges the importance of event order but ignores temporal information like duration. One common data processing technique in sequential analysis is to reduce the state-sequential data (events recorded in equal intervals) to event data (Bakeman & Gottman, 1997). This data representation only preserves the sequential aspect, disregarding the explicit reference to the duration. Although order and sequence can provide insights into the learning and collaboration process, temporal features such as position, duration, frequency, and rate also contain useful

information (Molenaar & Wise, 2022). For instance, duration may indicate levels of engagement (Nystrand et al., 2003; Zhou & Kang, 2022) or degrees of affect persistence (Baker et al., 2010). In addition, temporal patterns may not have a consistent distribution probability across different phases during group interactions (Fiore et al., 2010). For example, Kang et al. (2019) identified three distinct trajectories of group action similarity, providing empirical evidence that group exhibit varied similarity trends across different problem-solving phases. Thus, the lack of a complete account of temporal dimensions associated with data may limit the explanatory power and validity of the conclusions drawn. We, therefore, argue that temporal analysis needs to fully exploit the temporal information embedded in data by taking different dimensions of temporality into account.

The second practical issue is the focus on process data from a single modality such as discussion threads, verbal discussions, or log data (Lämsä et al., 2021). The lack of studies combining data from different sources limits the ability to interpret the temporal aspects of the collaboration process in CSCL, which is complex and multidimensional (Praharaj et al., 2021). Different actions like discourse, gesture, and visual attention are intertwined and contribute to sophisticated social interactions (Schneider et al., 2021), making a single modality inadequate to depict the whole picture. Recent research has illustrated the promise of multimodal data in terms of better prediction of learning outcomes and interpretation of complex processes (Cukurova et al., 2020). However, most methods either aggregated multimodal data or created discrete clusters of behaviours (Sharma & Giannakos, 2020). As such, limited attention has been paid to analyzing temporal patterns of multimodal features and understanding how these features mediate interactions from a process-oriented perspective.

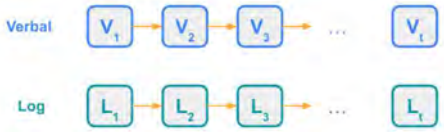
Alongside this, the third issue relates to the challenge of taking context into consideration. Contextual information can enrich the process data by enhancing the data storytelling and sequence interpretation (Martinez-Maldonado et al., 2019). This contextual knowledge can come from the same modality or different modalities. For example, when analyzing groups' verbal codes to understand how they co-construct knowledge, we can complement the discourse data with information from the same modality such as turn-taking or combine it with eye-gaze as a different modality. While most studies rely on manual and qualitative examination to make sense of the temporal patterns (Zhou et al., 2022), more sophisticated analytical approaches can contextualize these patterns for better interpretation.

The fourth issue pertains to determining the appropriate analytic unit to aggregate data. This unit or granularity defines the grain size of events that researchers code to create a meaningful unit for further analysis. Decision on the analytic unit has a profound impact on the patterns detected (Helske & Helske, 2021; Molenaar & Wise, 2022). Some data has clear and reliable boundaries to segment, such as discussion posts and learning sessions. In many other cases, however, researchers must make decisions on analytic units to aggregate or segment data without theoretical support. Furthermore, the appropriate granularity of time (size of time units) varies across modalities. For example, log data can be much finer grained than discourse data. Not only do researchers need to synchronize different modalities, but they also need to carefully decide the analytic unit that works for all modalities being analyzed.

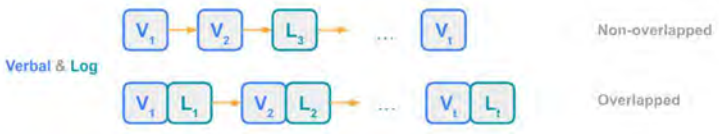
In summary, approaches taking both temporality and multimodality into account become particularly important for CSCL research. It is, however, challenging as the level of data complexity increases with new advances in multimodal technologies. In the following section, we give a brief review of current multimodal temporal analysis approaches. In particular, we discuss their affordances and limitations to addressing the issues discussed above.

2.2. Multimodal Temporal Analysis

Characterizing and analyzing the temporal patterns based on multimodal features is challenging yet important work in advancing the CSCL field. Several approaches have been proposed to examine multimodal data with the goal of understanding the temporal dynamics of the collaboration process. Figure 1 visualizes some examples of multimodal data representation as well as the temporal relationship being analyzed in existing approaches.

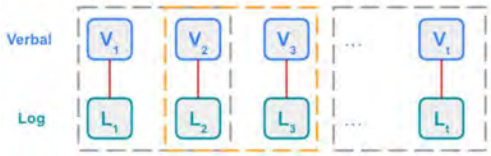


(a) An example of unimodal temporal analysis.



(b) An example of temporal analysis with non-overlapped and overlapped multimodal features.

ISSN 1929-7750 (online).



(c) An example of ENA (the red lines indicate no specification of relations direction).

Figure 1. Visualizations of multimodal feature representation and temporal relationships in different temporal analyses.
Note: The value of t indexes the analytic time unit.

Figure 1(a) represents the sequential analysis investigating the transitions of each modality independently. An alternative approach is to integrate multimodal features into one sequence, assuming that different features do not overlap as shown in the non-overlapped instance of Figure 1(b). This allows researchers to employ sequential analysis approaches to examine the transitions between these features or identify meaningful recurrent patterns within the sequence. For instance, Chang et al. (2017) collected discourse from chatrooms and activity logs from the web-based simulation to analyze the CPS process. These multimodal features were aligned chronologically and examined with lag sequential analysis to identify problem-solving patterns in successful and unsuccessful groups. This study presupposes that chatroom discourse and activity logs are non-overlapping events, which is applicable in web-based simulations where students are restricted to communicating through a chatroom. However, this assumption is rarely applicable in CSCL environments, especially in collocated contexts where students communicate and interact through multiple channels such as speech, eye contact, gesture, and physical actions (Martinez-Maldonado et al., 2019). As such, this approach is limited in its applications due to the strict assumption of the chronological alignment of multimodal features.

Another approach is encoding the unimodal sequence with contextual information collected from other sources (see the overlapped instance in Figure 1(b) where each feature incorporates information from different modalities within the same analytic time unit). This method of feature encoding was used to investigate relationships between discussions and physical actions and to identify temporal patterns in more and less collaborative groups with differential sequence mining (Martinez-Maldonado et al., 2013; Martinez et al., 2011). Although this approach demonstrated great potential to contextualize the unimodal sequence at a fine-grained level, it brings a practical challenge to statistical analysis. Incorporating additional information multiplies the number of possible sequences, leading to a substantial number of possible sequences and an increase in sequence variability. In addition, this approach can introduce rare events where certain encoded events only have a few instances. This results in the instability of summary statistics and decreased confidence in the conclusions drawn from statistics (Bakeman & Gottman, 1997). In other words, this encoded data representation may weaken the power to detect statistically significant differences and bias the frequency estimates, leading to a higher false discovery rate. Indeed, this issue was reflected in the choice of a larger p-value (i.e., $p = 0.1$) in the differential sequence mining in their follow-up study (Martinez-Maldonado et al., 2013). As a result, this representation requires a much larger dataset to make statistical inferences, which is not always available.

Instead of focusing on sequential order, Shaffer et al. (2009) introduced Epistemic Network Analysis (ENA) to quantify the co-occurrence between features and visualize their relationships in network models. This method was originally generated to analyze connections between cognitive and epistemic elements extracted from discourse data, it also holds the potential to apply to multimodal features with appropriate data representation. Figure 1(c) illustrates how ENA could analyze temporal co-occurrence between different modalities within the same and neighbouring analytic units, indicated by the dotted box. Recent works have combined ENA with temporal information to construct a trajectory (Brohinsky et al., 2021) or sequential analysis (Tan et al., 2022), allowing a more systemic account and accurate interpretation of temporal relations. This approach, however, falls short in identifying the lagging effects (Shaffer et al., 2009). Further, this approach counts the occurrences of adjacent events segment by segment (Csanadi et al., 2018). Decisions concerning the size of unit of analysis to segment data and model their relations within recent temporal context impact the results. For example, when analyzing data with reliable and natural boundaries like discourse or text data, researchers can use utterance as a meaningful analytic unit. However, the decision needs careful consideration and clear justification for other types of data.

2.3. Contribution and Research Objectives

These above-mentioned studies advanced the theoretical and methodological aspects of temporal analysis, particularly in CSCL environments. However, integrating multimodality into temporal analysis remains insufficiently explored. In response to this gap, we introduce a novel approach that serves as a methodological innovation in the field, focusing on the identification and analysis of temporal patterns within and across modalities. We utilize multilevel vector autoregression (mlVAR) to

contextualize a latent collaboration construct — joint attention — extracted from logs, alongside verbal interactions derived from video and audio sources. Figure 2 provides a visual representation of this multimodal temporal analysis, facilitating the examination of dependent temporal relations between multimodal data. This approach allows us to estimate both temporal and contemporaneous correlations between multimodal features nested in groups. A previous study has applied this approach to analyze qualitative codes of verbal communications during a CPS task and reported the temporal networks (i.e., sequential relations between different time points) without reporting contemporaneous networks (i.e., concurrent relations within the same time point; Zhou et al., 2022). We, therefore, extend this method to analyze features extracted from multiple data sources and report both temporal and contemporaneous networks to provide a more comprehensive description of the collaborative learning processes.

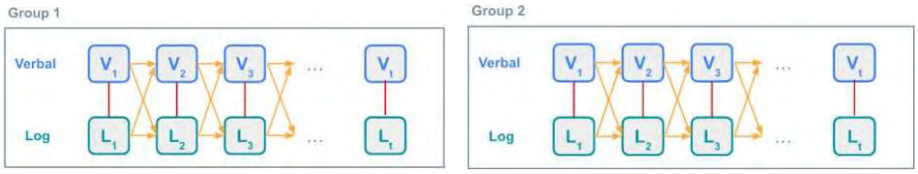


Figure 2. Visualization of multimodal feature representation and temporal relationships in mlVAR.

We argue that this methodological advancement enriches the existing literature on temporal analysis three-fold. First, it shows the potential of enhancing unimodal data representation by integrating contextual information, enabling a more nuanced model to capture student behaviours. Second, it supports both contemporaneous and temporal (lagging) analysis, providing a more holistic picture to describe temporal relationships within and between features. Third, it broadens the scope of temporal analysis, which has traditionally investigated CSCL through the lens of sequential order (Sarmiento & Stahl, 2008; Zheng et al., 2019). This work introduces an approach that probes temporality both in terms of sequence and duration. We demonstrate the potential of mlVAR to explore behaviour patterns, with a focus on identifying feature co-occurrence and transitions that differentiate low- and high-achieving groups. This paper, therefore, primarily aims to do two things: 1) showcase the capabilities of mlVAR in multimodal temporal analysis and 2) underscore the importance of considering both dimensions of temporality.

3. Methods

3.1. CEASAR

Our learning platform is an immersive astronomy simulation designed to empower students to observe and explore the night sky across space and time by modifying locations on Earth and time within the system. This simulation provides access to a star and constellation database from three different perspectives: Horizon, Star, and Earth. Figure 3 illustrates screen captures of tablet and AR headset users, where Horizon view (left) provides a first-person view of the sky, Star view (middle) removes the horizontal limitation and gives full access to the celestial sphere, and Earth view (right) places the student in orbit. To facilitate small group collaboration, this platform allows group members to share their perspectives, annotations, and highlighted stars or constellations across multiple devices (tablet and Microsoft HoloLens2 AR headset) in real time.

3.2. Participants and Tasks

The participants included 77 undergraduates enrolled in an introductory astronomy course at a mid-western university in the United States. To familiarize students with the platform, two introductory sessions were conducted to introduce the simulation and its functions in both tablets and AR headsets. The following week, students were required to collaboratively complete a task called “Lost at Sea” during a 50-minute lab session. Each group of three to four students (a total of 25 groups) had one AR headset and two touch-based tablets. The task required groups to leverage their knowledge to determine the longitude and latitude of an unknown location randomly assigned to each group within the simulation. Three subtasks were designed to help solve the assignment step-by-step: 1) determining the hemisphere, 2) identifying navigational stars or constellations, and 3) calculating the longitude and latitude. Aside from the group task, students were required to complete individual pre- and post-paper-based assessments before and after the task session. Each assessment took about five minutes.

Figure 3. CEASAR screen captures: (left) Horizon scene in a tablet, (centre) Star scene in a tablet, (right) Earth scene in AR.

3.3. Group Exclusion and Classification

In the qualitative coding process, nine groups were removed due to data unavailability and insufficient group members, resulting in 16 groups being further analyzed. This study used pre- and post-assessments to quantitively measure learning gains. These assessments contained one open-ended question to evaluate the change in student knowledge of latitude and longitude calculation. Student responses were scored on a scale of 0 to 2, reflecting their understanding of astronomy concepts and phenomena. A score of 0 was assigned if the response did not contain any key astronomy concepts, a score of 1 for responses that included only fundamental concepts, and a score of 2 for answers that contained all desired astronomy concepts. We then computed individual normalized gains (i.e., post – pre / post-max – pre) to obtain average normalized learning gains for each group. Upon examining the distribution of group learning gains, we identified a gap between 0.1667 and 0.3125, which potentially indicates a divergence in knowledge gain. We thus classified 16 groups into low-achieving (n = 7) with a range [–0.2222, 0.1667] and high-achieving groups (n = 9) with a range [0.3125, 0.6875]. This classification ensured a relatively balanced distribution of groups and a sufficient gap in knowledge gains between the two performance categories.

3.4. Data Source and Features

This study collected video recordings, screen capture, and log data from the simulation platforms for tablet and AR headsets. Interaction logs were the primary data source used to identify joint attention within simulation across devices. Logs were recorded as rows of events, where event = {Username, Groupname, Device, Activity, Event, UTC time, Heading vectors, Simulation time, Crashsite, Location, Scene, Selected object, Selected star}. A new event was generated each time students moved their devices to change the direction of view, selected a star or constellation, chose a different scene, or manipulated the simulation time within the platform. We included video recordings as the secondary data source to analyze verbal interactions and screen capture to synchronize different data sources.

3.4.1. Joint Attention

Joint attention (JA) was operationalized as the moment when students shared a common focus on the simulated objects across different digital devices. Since our simulation platform allows students to explore areas in the same scene and switch between different scenes (Earth, Star, and Horizon), JA under different scenarios may not indicate the same level of collaboration and shared attention. For example, students could switch to the Earth scene together (see Figure 3-right) while they may not necessarily look at the same area of the Earth. Another scenario is screen overlapping, which requires students to adjust their screens to observe the same constellation in Horizon (see Figure 3-left). This level of visual synchronization requires extra effort to move headsets or screens to locate the targeted object. More importantly, students may have made "attention contact" with each other (Siposova & Carpenter, 2019), where students intentionally communicate to align screens and have awareness of each other's focus. Screen overlapping thus reflects a conscious decision to achieve joint attention.

Table 1. Feature Descriptions

Feature	Description
JA state I	no student triggers any event within the 20 seconds
JA state II	students explore in different scenes OR only one student is active
JA state III	both students stay in the Earth or Star scene
JA state IV	both students stay in the Horizon scene but without screen overlaps
JA state V	both students stay in the Horizon scene and have screen overlaps
Verbal Interaction	group has conceptual discussions or task-relevant discussions

Note: The 20-second inactivity gap is inspired by Martinez et al. (2011) and informed by in-class observations.

To capture these various degrees of joint attention, we developed a framework shown in Table 1. These JA states were extracted from logs; that is, based on scene and field-of-view information, we computed percentages of screen overlap between two devices (Diederich et al., 2021; Zhou & Kang, 2022). Unlike existing studies that dichotomized JA into a binary variable (e.g., whether or not groups have JA moments), our operationalization highlights five levels of JA varying in degree of visual synchronization. As described in Table 1, higher levels, such as JA states IV or V, require more effort and communication to achieve, thus indicating an advanced level of attention coordination to build a shared problem space and facilitate group sense-making. The initial coding generated a sequence of JA states for each dyad within each group, resulting in three device pairs (i.e., tablet1–tablet2, tablet1–AR, and tablet2–AR). We noticed a tendency for groups to use only two devices most of the time, suggesting that examining all dyad sequences may lead to inaccurate conclusions due to inconsistent device usage. Therefore, one dyad sequence was selected to represent the group holistically based on the level of involvement.

3.4.2. Verbal Interaction

To examine the interplay between JA states and verbal interactions, we transformed qualitative codes into a second-level binary variable (i.e., indicating the presence or absence of conceptual discussions). These codes were derived from a turn-by-turn coding process using a scheme developed by our team (Planey et al., 2023), inspired by Mercier et al. (2017). We focused on task-relevant discussions between group members. After co-coding one group's data, two researchers coded 20% of the remaining data and achieved an agreement rate of at least 80% across all categories.

3.5. Multimodal Data Representation

We simplified the multimodal matrix developed by Echeverria et al. (2019) to organize our features shown in Figure 4. Each second-level feature could be seen as a *multimodal observation* associated with one specific *dimension of collaboration* (e.g., attentional, verbal). We aggregated these observations by *segment*, the smallest analytic time unit. Each segment represents a fixed unit of 30 seconds, starting from the moment when groups began to work on the first task. We tested different unit sizes to find an appropriate segment size to balance the different granularities of JA state and verbal discussion. Table 2 shows the average duration of each JA state and the round of discussion (the ending point of each discussion was defined as no discussion for 10 seconds). A large analytic unit could capture the co-occurrence of different features but may not be sufficiently sensitive to JA state transition. Conversely, a small analytic unit may capture more transitions but has the issue of dominating self-transition that may suppress non-self-transitions (Karumbaiah et al., 2018). To accommodate these two modalities, we set the range of analytic units between 20 and 40 seconds to capture transitions of JA states while avoiding dominating self-transitions of discussion. The 30-second interval was eventually selected since it yielded smaller residuals and was the best fit for the model. As a robust check, we examined 20 and 40 seconds and found the resulting patterns were comparable.

Table 2. Average Duration of JA and Verbal Features (in Seconds)

	JA state I	JA state II	JA state III	JA state IV	JA state V	Discussion
Low	38.56	44.18	25.50	26.14	24.53	92.63
High	53.21	37.68	20.57	34.37	53.83	101.15

The duration of each feature was transferred into a binary variable (i.e., 0 or 1), indicating the presence of this *dimension of collaboration* (e.g., JA, verbal interaction) in the *segment*. We chose the threshold of 10 seconds (half of the smallest average duration of all features) to decide whether this feature occurred during the analytic time unit. If the duration of a feature was less than 10 seconds, it was labelled as zero (0). This value was chosen to account for the extreme case where one feature spanned two units with equal proportion; otherwise, this feature would only be recorded in the unit with a larger proportion. This allowed for different features to be observed in the same analytic unit. For example, both JA I and JA II states were documented in the second analytic unit as shown in Figure 4. We dichotomized all features rather than using the duration value as our research focuses more on the co-occurrence and temporal relations between features (e.g., whether higher level JA state co-occurs with conceptual discussion or whether conceptual discussion leads to specific JA states). The duration or persistence could be inferred indirectly by self-transition likelihood: a feature with stronger self-transition indicates longer persistence. In this way, we could obtain information about persistence without adding complexity to the interpretation of temporal associations. For instance, a stronger temporal relation could only be interpreted as a higher likelihood of occurrence rather than a feature with a longer duration.

Dimensions of Collaboration

Index		Attentional					Verbal
Group	Analytic unit	JA state I	JA state II	JA state III	JA state IV	JA state V	Discussion
						Multimodal	Observations
3	1	1	0	0	0	0	1
3	2	1	1	0	0	0	1 Segment
...							
3	53	0	0	0	0	1	0

Figure 4. Multimodal matrix.

3.6. Analysis Method

We utilized multilevel vector autoregression (mlVAR), which is a combination of vector autoregression (VAR) and multilevel modelling (Bringmann et al., 2013). The autoregressive (AR) model predicts each variable at time point *t* with the same variable from previous time points *t* – *h* (i.e., h is called time lag). VAR is a multivariate extension of the AR model, which means it models the time dynamics by predicting one variable at time point *t* with all variables from time points *t* – *h*. By combing VAR and multilevel modelling, we can further estimate temporal associations between multiple variables under a multilevel framework (e.g., when measurements are nested within each subject, students are nested within groups) described by the following equation where *i* represents the group index:

$$\text{JA state I}_t^i = {\beta_0}^i + \beta_1 \text{JA state I}_{t-1}^i + \beta_2 \text{JA state II}_{t-1}^i + \beta_3 \text{JA state III}_{t-1}^i + \ldots + \beta_6 \text{Discussion}_{t-1}^i + \epsilon_{t-1}^i$$

We used the mlVAR package (Epskamp et al., 2021) in R to build multiple regression models for each variable (i.e., each type of multimodal observations) and estimate the parameters; for example,

$$\text{JA state I}_t^i = {\beta_{01}}^i + \beta_{11} \text{JA state I}_{t-1}^i + \beta_{21} \text{JA state II}_{t-1}^i + \beta_{31} \text{JA state III}_{t-1}^i + \ldots + \beta_{61} \text{Discussion}_{t-1}^i + \epsilon_{t-1;1}^i$$

$$\text{JA state II}_t^i = {\beta_{02}}^i + {\beta_{12}}^i \text{JA state I}_{t-1}^i + {\beta_{22}}^i \text{JA state II}_{t-1}^i + {\beta_{32}}^i \text{JA state III}_{t-1}^i + \cdots + {\beta_{62}}^i \text{Discussion}_{t-1}^i + \epsilon_{t-1;2}^i$$

…

$$\text{Discussion}_t^i = {\beta_{06}}^i + {\beta_{16}}^i \text{JA state I}_{t-1}^i + {\beta_{26}}^i \text{JA state II}_{t-1}^i + {\beta_{36}}^i \text{JA state III}_{t-1}^i + \cdots + {\beta_{66}}^i \text{Discussion}_{t-1}^i + \epsilon_{t-1;3}^i$$

Following a two-step estimation, this modelling approach estimates the temporal, contemporaneous, and between-subject networks. In the first step, mlVAR models the temporal associations shown in the equations above for each subject, which refers to each group in our case. Each variable (i.e., JA state or discussion) is predicted by all variables, including itself at a previous time point *t* – *h*. Since predictors also include the previous value of this variable itself (e.g., use JA state I_{t-1}^i to predict JA state I_t^i), we can also observe self-transition in the network. Aside from temporal correlations, mlVAR estimates between-subjects associations based on the random intercept β_0. In the second step, residuals from the previous model were used to build a sequential univariate multilevel regression model. As such, this model predicts residuals of one variable by residuals of all other variables within the same time frame, resulting in the estimation of partial contemporaneous correlations (Epskamp et al., 2018).

Two assumptions need to hold to utilize mlVAR: 1) equally spaced time points, which means fixed analytic time unit to aggregate data, and 2) stationarity, which means the mean and variance do not change as a function of time. We first checked the stationarity assumption using Augmented Dickey-Fuller (Mushtaq, 2011) and the Kwiatkowski-Phillips-Schmidt-Shin test (Syczewska, 2010). Both tests indicated no trend in the data ($p < 0.01$). This means a constant autocorrelation structure where the strength of temporal dependence remains the same across various time points. In addition, the assumption of equally spaced time points was satisfied since we used a fixed analytic time unit to aggregate data. Then we applied the mlVAR to model temporal and contemporaneous associations for low- and high-achieving groups, respectively. In this current paper, we only reported the results from temporal and contemporaneous networks and ignored between-subjects networks, as how each group differs is not our focus, and we do not have sufficient subjects (i.e., the number of groups in our case) to test the between-subject effects as well.

4. Results

4.1. Temporal Network

The inferred temporal networks are presented in Figure 5. Each node represents one of the features. Weighted arrows represent temporal relations between the features. This relationship can be negative, as suggested by the red arrow in visualization, or positive, as suggested by the green arrow. In addition, the estimated parameters of lagged predictors are translated into the thickness of the arrows: a thicker arrow between two nodes represents a stronger relation between the predictor at time point $t - 1$ and the predicted variable at time point t.

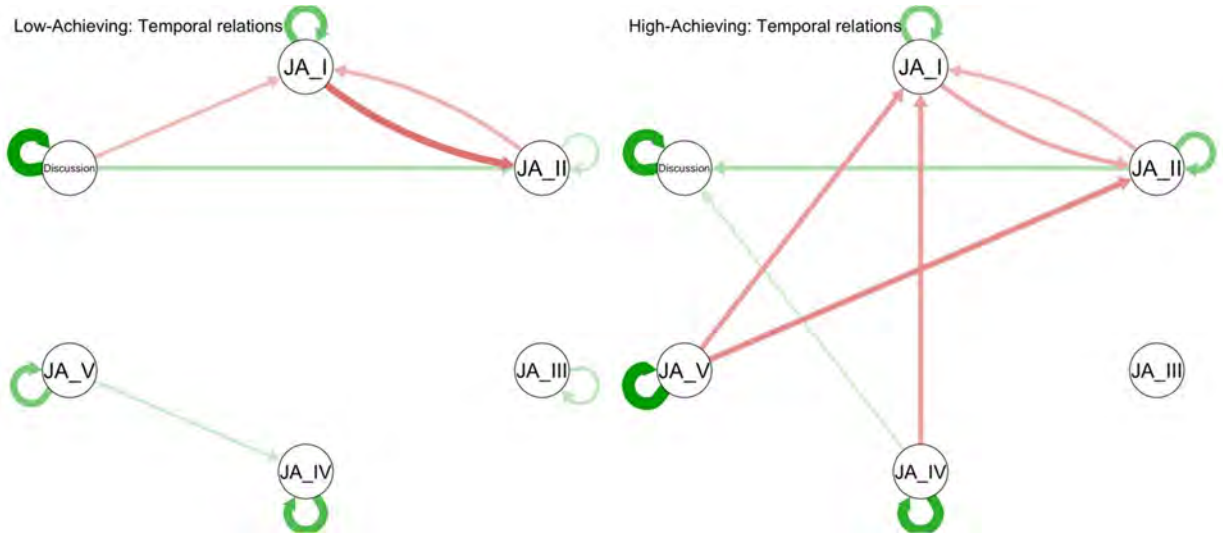


Figure 5. Temporal Networks of low-achieving (left) and high-achieving (right) groups.
Note: This network only visualizes arrows that surpass the threshold of significance (i.e., $p < 0.05$).

As shown in Figure 5, the comparison of temporal networks between low- and high-achieving groups yields several interesting findings. First, it is expected that both high- and low-achieving groups exhibit positive self-transitions. This is due to the deliberate selection of the analytic unit size smaller than the average length of each feature, as discussed in section 3.5, to capture both self- and non-self-transitions. As a result, we can infer persistence based on the strength of the self-transition arrow. Although both groups have positive self-transitions of the highest level of joint attention (i.e., JA V), high-achieving groups show a much stronger transition from the previous JA V to the current one ($\beta_{high} = 0.558$, $p < 0.05$; $\beta_{low} = 0.310$, $p < 0.05$). This suggests that each time high-achieving groups achieve screen overlapping (JA V), they stay in this state longer than low-achieving groups. Such sustained engagement in joint attention may facilitate a shared understanding of the problem space. This provides necessary time for in-depth and more thorough discussion, potentially leading to successful knowledge co-construction. This calls for future studies of examining the associations between such behaviours and groups' knowledge co-construction. This finding pinpoints the necessity of taking different temporal dimensions (i.e., duration) into account. Aggregating data in total or treating them as event sequences without the consideration of duration will fail to capture such information.

Another finding relevant to JA V is how this feature predicts the subsequent features. For high-achieving groups, JA V is a significant negative predictor of JA I and JA II (fixed effect β are –0.211, –0.242 respectively with $p < 0.05$). This result suggests that after leaving the state of screen overlapping in Horizon, high-achieving groups are less likely to enter the two other JA states, which may indicate a group's low-level engagement and attention coordination with their devices. We did not find any positive temporal relations starting from JA V to other states. One potential reason is the strong autoregressive effect of JA V (i.e., self-transition), which may overshadow the influence of other predictors. This suggests a potential future avenue for using a larger analytic unit size or temporal lags, which eliminates self-transition cases and thus explores what distinct JA or verbal features follow JA V.

Lastly, we found an interesting difference between JA states and conceptual discussion. For low-achieving groups, none of the JA states was identified as a significant predictor of discussion; that is, no significant association between joint attention on the device and conceptual discussion. This means that these groups rarely initiated conversations around collaborative explorations on their devices. In contrast, two JA states were found to predict conceptual discussion for high-achieving groups

in the following time frame: JA states II and IV. Both states represent individual explorations without screen overlapping. These two transitions (JA state II → Discussion; JA state IV → Discussion) indicate that individual explorations in the simulation actually led to conceptual discussions, which results in successful CPS for high-achieving groups. Individual explorations in high-performing groups seem to contribute to knowledge co-construction through initiating and inviting productive discussions, which was not observed in low-achieving groups. This key difference aligns with empirical studies on individual contributions within the group (Chiu, 2000) and further confirms the importance of sustaining group engagement in socially shared processes and collaborative attempts at joint meaning-making (Volet et al., 2017).

4.2. Contemporaneous Network

mlVAR also provides information about co-occurrence relations within the same analytic time unit: contemporaneous relations. Figure 6 shows the estimated contemporaneous networks for two learning performance groups. This network represents correlations between nodes within the same time frame, partialling out the temporal effects discussed above. All the correlations between JA states are negative, suggested by the red colour in Figure 6. It was not surprising since they belong to the same collaboration dimension, making them less likely to co-occur in the same time window. Thus, focusing on the association between JA states and discussion is more meaningful. Comparing these two networks, the main difference is that only high-achieving groups show a significant negative correlation between discussion and JA state I. This negative association indicates a lower probability of co-occurrence between JA state I and discussion. Since JA state I was identified when none of the group members stayed active on their devices, this association means when high-achieving groups had conceptual discussions, they were less likely to be inactive on their devices and tended to engage with the simulation platform. This finding further indicates that high-achieving groups engage in both physical and virtual spaces through verbal communication and digital interactions.

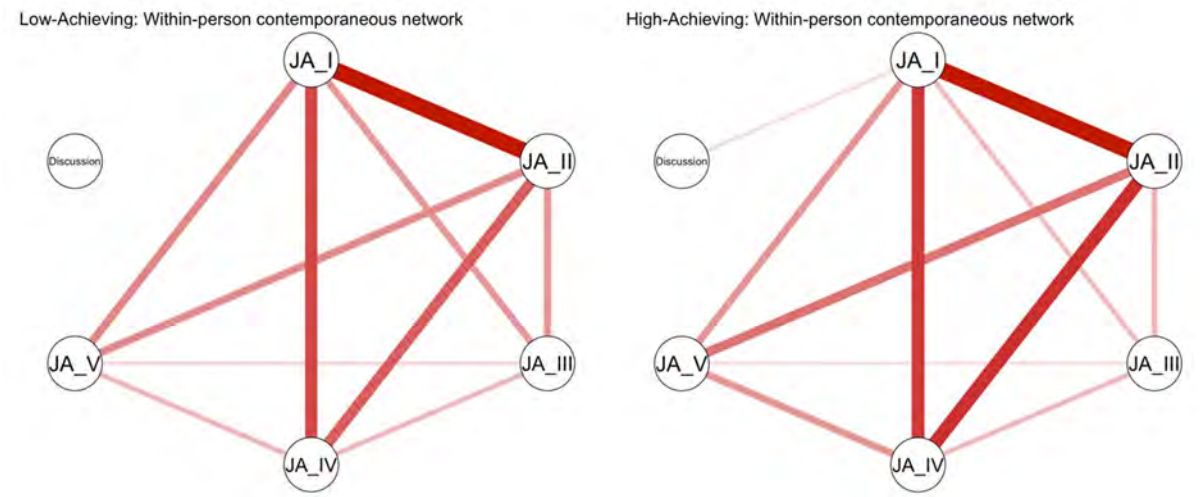


Figure 6. Contemporaneous networks of low-achieving (left) and high-achieving (right) groups.
Note: This network only visualizes arrows that surpass the threshold of significance (i.e., $p < 0.05$).

5. Discussion

Temporal analysis of process data is gaining prominence in the field of CSCL. However, most existing research has examined temporal patterns from a single modality. This narrow scope limits pattern interpretation and lacks contextual information. While some studies began to combine multimodal data to enrich current understanding, they typically take an effect-oriented approach (Janssen et al., 2010), focusing on predicting outcomes rather than illustrating intermodal dynamics (Sharma & Giannakos, 2020). Such methods often fail to show how different modalities interact and contribute to group interactions from a process-oriented perspective (Dillenbourg et al., 1996). To tackle these challenges, we present an empirical example of multimodal temporal analysis that applied mlVAR analysis to model and visualize collaborative interactions. Specifically, our introduced method did the following: 1) combined multimodal features extracted from the log and video data, 2) contextualized the joint attention with verbal interactions, 3) conceptualized temporality in terms of both time passage and sequential order, 4) analyzed temporal and contemporaneous relationships between features, and 5) conducted a multilevel analysis to account for within-group interdependencies.

We argue that this approach, combining temporal and multimodal analysis, has important implications for collaboration analytics. First, it provides empirical support for the benefits of incorporating multiple modalities when interpreting unimodal process data (Cukurova et al., 2020). Building upon previous findings (Zhou & Kang, 2022) on joint attention, this paper contextualizes it with the information regarding whether groups engaged in conceptual discussions, providing complementary insights into how groups interact beyond the digital sphere. Our results observed that high-achieving groups were more likely to initiate discussions after individual exploration (i.e., JA state II → Discussion; JA state IV → Discussion), signifying a critical moment where individual discovery is transferred to collective knowledge. This echoes the notion of timely joint attention (Barron, 2003). Although attentional engagement is a prerequisite for successful collaboration, joint attention does not need to be always maintained. Instead, it should be regained at solution-critical times (Teasley & Roschelle, 1993), those key moments when collective focus is required on problem resolution and shared attention has a particularly important impact. This observation implies that individual explorations could benefit the entire group by seeking and providing information to facilitate the exchange of ideas when groups re-establish shared attention. This supports that idea divergence can contribute to collective convergence (Kapur et al., 2008). The incorporation of a new modality identifies a key moment in CPS, explaining when and how joint attention facilitates productive collaboration. This insight also carries implications for scaffolding in open-ended problem-solving scenarios: it may be more effective to provide support during solution-critical moments, particularly during transitions from individual exploration to collective discussion.

Second, this work exemplifies the value of adopting a process-oriented perspective to unpack collaboration. By examining sequential and contemporaneous associations between multimodal features, our findings indicate that high-achieving groups demonstrate more temporal patterns, such as sustained high-level joint attention and co-occurrence between engagement with the simulation and conceptual discussion. This difference can be attributed to the ability to coordinate and maintain shared attention and initiate discussions in a productive way that fosters the group's collaborative problem-solving process. This connection implies an important role of screen overlapping in knowledge co-construction; students actively coordinate digital screens to create a shared problem space, grounding discussions and building upon each other's findings (Stahl, 2006). This work highlights the importance of temporal accounts, as aggregating all modalities and ignoring their interconnections would not reveal this insight. Future studies could apply this approach to investigate how different modalities like gestures, physical actions, discourse, and eye-gaze are intertwined to orchestrate collaboration processes and determine when support is necessary.

Third, this work underscores the importance of conceptualizing temporality both as the passage of time and sequential order in time (Molenaar & Wise, 2022) and investigating different temporal features in practice. While approaches like sequential analysis provide valuable insights into the chronological order of learning activities and patterns of student behaviours, they tend to only emphasize sequential aspects of events without an explicit reference to the duration of events (Lämsä et al., 2021). This study leverages the multimodal matrix to record both dimensions of temporality (i.e., duration and order). By coding the occurrence of each feature in each analytic time unit, we can infer the duration through self-transitions and detect sequential arrangement by investigating their transition patterns. This inferential statistic not only reveals descriptive information (such as the relatively short or long duration of event passage) but also allows for statistical comparison within features (whether this feature has stronger self-transitions or non-self-transitions) and between groups (whether low- and high-achieving groups exhibit statistically different self-transition patterns that indicate the distinguishing power of duration). Our results illustrate that low- and high-achieving groups differ in both dimensions: 1) duration, as high-achieving groups show a more persistent JA state V, and 2) order, as suggested by distinct transition patterns between learning performance groups discussed above. This finding also provides empirical evidence for the hypothesis in our preliminary study (Zhou & Kang, 2022) that a longer JA state may indicate a higher level of engagement and collaboration. Overall, our results show that both dimensions can yield meaningful insights and support distinct claims regarding the temporal characteristic of the constructs.

Last, this work provides researchers with a reusable workflow for multimodal data preparation and representation. This data representation can incorporate additional modalities like gestures or gaze by converting raw data into sequential categorical data within fixed analytic units. The resulting mlVAR networks enable statistical comparisons to identify temporal and contemporaneous differences across experimental conditions or groups. This work also highlights best practices for interpreting mlVAR outputs, including examining self-transitions (duration), cross-feature transitions (sequence), and contemporaneous connections within the same analytical window.

In summary, mlVAR shows promise for a comprehensive assessment of temporal dynamics and contemporaneous interactions between multiple data sources. By detailing the end-to-end workflow and demonstrating analytical value, this paper seeks to help fellow researchers apply this technique to enrich unimodal data and uncover the temporal interplay of multimodal constructs.

5.1. Limitations and Future Research

This study has several limitations that should be addressed in future studies. One limitation is the small sample size containing only 16 groups. In addition, we made some parameter decisions that may impact the results, including the selection of the analytic time unit and coding thresholds. Third, we chose a small analytic unit to investigate the persistence of self-transitions, resulting in weaker power to detect sequential associations between different features (Karumbaiah et al., 2018). As such, many parameter decisions were shaped by our specific research contexts with a limited theoretical basis. Readers should be aware that altering these parameters could significantly influence the patterns identified when applying this approach. There is a need for future research to provide stronger theoretical and empirical guidance on selecting appropriate parameter values. We plan to conduct sensitivity analyses on future work to systematically evaluate the robustness of findings across different parameter settings. Overall, these limitations highlight the preliminary nature of this work, requiring further validation on an expanded dataset.

6. Conclusion

Collaboration is a complex multidimensional process. There is increasing attention toward measuring and modelling collaboration with fine-grained process data from multimodal sensors. However, how these data intersect and mediate the group interaction remains underexplored. This paper proposes a multimodal temporal approach to unpack the intertwined relationship between different modalities during CPS in an immersive learning environment. Specifically, we utilized mlVAR to analyze the temporal and contemporaneous associations between a latent feature (joint attention) and an observable feature (verbal discussion). Distinct behaviour patterns identified in high-achieving groups suggest that they tend to initiate discussion after individual explorations within the simulation. This indicates a key moment in CPS to transfer individual discovery to collective understanding through discussion. Furthermore, our result illustrates the necessity of a more complete account of temporality (both the passage of time and sequential order) as low- and high-achieving groups show significant differences in both dimensions. In summary, this work showcases the benefits of mlVAR for process-oriented investigation of temporal and multimodal interrelationships. By detailing the necessary data structuring and analytical procedures, this paper aims to assist other MMLA researchers in productively adopting mlVAR to unpack new insights into collaborative learning dynamics from multimodal data.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The publication of this article received financial support from the National Science Foundation under grant #1822796.

Acknowledgement

We sincerely thank the entire CEASAR team and the Concord Consortium for their support. Special thanks to Luc Paquette for his insightful feedback on visualization and conceptualization.

References

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511527685

Baker, R. S. J. d., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human–Computer Studies*, *68*(4), 223–241. https://doi.org/10.1016/j.ijhcs.2009.12.003

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

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE*, *8*(4), e60188. https://doi.org/10.1371/journal.pone.0060188

Brohinsky, J., Marquart, C., Wang, J., Ruis, A. R., & Shaffer, D. W. (2021). Trajectories in epistemic network analysis. In A. R. Ruis & S. B. Lee (Eds.), *Advances in quantitative ethnography* (pp. 106–121). Springer. https://doi.org/10.1007/978-3-030-67788-6_8

Chang, C.-J., Chang, M.-H., Chiu, B.-C., Liu, C.-C., Fan Chiang, S.-H., Wen, C.-T., Hwang, F.-K., Wu, Y.-T., Chao, P.-Y., Lai, C.-H., Wu, S.-W., Chang, C.-K., & Chen, W. (2017). An analysis of student collaborative problem solving activities mediated by collaborative simulations. *Computers & Education*, *114*, 222–235. https://doi.org/10.1016/j.compedu.2017.07.008

Chango, W., Lara, J. A., Cerezo, R., & Romero, C. (2022). A review on data fusion in multimodal learning analytics and educational data mining. *WIREs Data Mining and Knowledge Discovery*, *12*(4), e1458. https://doi.org/10.1002/widm.1458

Chen, B., Knight, S., & Wise, A. F. (2018). Critical issues in designing and implementing temporal analytics. *Journal of Learning Analytics*, *5*(1). https://doi.org/10.18608/jla.2018.53.1

Chen, B., Resendes, M., Chai, C. S., & Hong, H.-Y. (2017). Two tales of time: Uncovering the significance of sequential patterns among contribution types in knowledge-building discourse. *Interactive Learning Environments*, *25*(2), 162–175. https://doi.org/10.1080/10494820.2016.1276081

Chiu, M. M. (2000). Group problem-solving processes: Social interactions and individual actions. *Journal for the Theory of Social Behaviour*, *30*(1), 26–49. https://doi.org/10.1111/1468-5914.00118

Csanadi, A., Eagan, B., Kollar, I., Shaffer, D. W., & Fischer, F. (2018). When coding-and-counting is not enough: Using epistemic network analysis (ENA) to analyze verbal data in CSCL research. *International Journal of Computer-Supported Collaborative Learning*, *13*(4), 419–438. https://doi.org/10.1007/s11412-018-9292-z

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

Diederich, M., Kang, J., Kim, T., & Lindgren, R. (2021). Developing an in-application shared view metric to capture collaborative learning in a multi-platform astronomy simulation. *Proceedings of the 11th International Conference on Learning Analytics and Knowledge* (LAK '21), 12–16 April 2021, Irvine, CA, USA (pp. 173–183). ACM Press. https://doi.org/10.1145/3448139.3448156

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. http://infoscience.epfl.ch/record/33761

Dillenbourg, P., Järvelä, S., & Fischer, F. (2009). *The evolution of research on computer-supported collaborative learning: From design to orchestration* (pp. 3–19). https://doi.org/10.1007/978-1-4020-9827-7_1

Echeverria, V., Martinez-Maldonado, R., & Buckingham Shum, S. (2019). Towards collaboration translucence: Giving meaning to multimodal group data. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 4–9 May 2019, Glasgow, Scotland, UK (Paper no. 39). ACM Press. https://doi.org/10.1145/3290605.3300269

Emerson, A., Cloude, E. B., Azevedo, R., & Lester, J. (2020). Multimodal learning analytics for game-based learning. *British Journal of Educational Technology, 51*(5), 1505–1526. https://doi.org/10.1111/bjet.12992

Epskamp, S., Deserno, M. K., Bringmann, L. F., & Veenman, M. (2021). *mlVAR: Multi-level vector autoregression* (0.5) [Computer software]. https://CRAN.R-project.org/package=mlVAR

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The Gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, *53*(4), 453–480. https://doi.org/10.1080/00273171.2018.1454823

Fiore, S. M., Rosen, M. A., Smith-Jentsch, K. A., Salas, E., Letsky, M., & Warner, N. (2010). Toward an understanding of macrocognition in teams: Predicting processes in complex collaborative contexts. *Human Factors*, *52*(2), 203–224. https://doi.org/10.1177/0018720810369807

Helske, J., & Helske, S. (2021). seqHMM: Mixture hidden Markov models for social sequence data and other multivariate, multichannel categorical time series (1.2.0) [Computer software]. https://CRAN.R-project.org/package=seqHMM

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*(2), 139–154. https://doi.org/10.1007/s10648-010-9131-x

Kang, J., An, D., Yan, L., & Liu, M. (2019). Collaborative problem-solving process in a science serious game: Exploring group action similarity trajectory. In C. F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.), *Proceedings of the 12th International Conference on Educational Data Mining* (EDM2019), 2–5 July 2019, Montréal, Quebec, Canada (pp. 336–341). International Educational Data Mining Society.

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. https://doi.org/10.1016/j.compedu.2007.04.007

Karumbaiah, S., Andres, J. M. A. L., Botelho, A. F., Baker, R. S., & Ocumpaugh, J. S. (2018). The implications of a subtle difference in the calculation of affect dynamics. In J. C. Yang, M. Chang, L.-H. Wong, & M. T. Rodrigo (Eds.), *Proceedings of the 26th International Conference on Computers in Education* (ICCE 2018), 26–30 November 2018, Manila, Philippines (pp. 29–38). Asia-Pacific Society for Computers in Education. http://icce2018.ateneo.edu/wp-content/uploads/2018/12/C1-04.pdf

Knight, S., Wise, A. F., & Chen, B. (2017). Time for change: Why learning analytics needs temporal analysis. *Journal of Learning Analytics*, *4*(3), 7–17. https://doi.org/10.18608/jla.2017.43.2

Lämsä, J., Hämäläinen, R., Koskinen, P., Viiri, J., & Lampi, E. (2021). What do we do when we analyse the temporal aspects of computer-supported collaborative learning? A systematic literature review. *Educational Research Review*, *33*, 100387. https://doi.org/10.1016/j.edurev.2021.100387

Martinez, R., Yacef, K., Kay, J., Al-Qaraghuli, A., & Kharrufa, A. (2011). Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop. In M. Pechenizkiy et al. (Eds.), *Proceedings of the 4th Annual Conference on Educational Data Mining* (EDM2011), 6–8 July 2011, Eindhoven, Netherlands (pp. 111–120). International Educational Data Mining Society. https://educationaldatamining.org/EDM2011/wp-content/uploads/proc/edm11_proceedings.pdf

Martinez-Maldonado, R., Dimitriadis, Y., Martinez-Monés, A., Kay, J., & Yacef, K. (2013). Capturing and analyzing verbal and physical collaborative learning interactions at an enriched interactive tabletop. *International Journal of Computer-Supported Collaborative Learning*, *8*(4), 455–485. https://doi.org/10.1007/s11412-013-9184-1

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

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. https://doi.org/10.1111/bjet.12351

Molenaar, I., & Wise, A. F. (2022). Temporal aspects of learning analytics: Grounding analyses in concepts of time. In C. Lang, G. Siemens, A. F. Wise, D. Gašević, & A. Merceron (Eds.), *Handbook of learning analytics* (pp. 66–76). https://doi.org/10.18608/hla22.007

Mushtaq, R. (2011, August 17). Augmented Dickey Fuller test. *Econometrics.* https://doi.org/10.2139/ssrn.1911068

Nizam, S. S. M., Zainal Abidin, R., Che Hashim, N., Lam, M. C., Arshad, H., & Abd Majid, N. A. (2018). A review of multimodal interaction technique in augmented reality environment. *International Journal on Advanced Science, Engineering and Information Technology*, *8*(4–2), 1460–1469. https://doi.org/10.18517/ijaseit.8.4-2.6824

Noroozi, O., Alikhani, I., Järvelä, S., Kirschner, P. A., Juuso, I., & Seppänen, T. (2019). Multimodal data to design visual learning analytics for understanding regulation of learning. *Computers in Human Behavior*, *100*, 298–304. https://doi.org/10.1016/j.chb.2018.12.019

Nystrand, M., Wu, L. L., Gamoran, A., Zeiser, S., & Long, D. A. (2003). Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse Processes*, *35*(2), 135–198. https://doi.org/10.1207/S15326950DP3502_3

Phon, D. N. E., Ali, M. B., & Halim, N. D. A. (2014). Collaborative augmented reality in education: A review. *2014 International Conference on Teaching and Learning in Computing and Engineering*, 11–13 April 2014, Kuching, Malaysia (pp. 78–83). IEEE Computer Society. https://doi.org/10.1109/LaTiCE.2014.23

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. https://doi.org/10.1007/s11412-023-09398-w

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. https://doi.org/10.1109/TLT.2021.3097766

Reimann, P. (2009). Time is precious: Variable- and event-centred approaches to process analysis in CSCL research. *International Journal of Computer-Supported Collaborative Learning*, *4*(3), 239–257. https://doi.org/10.1007/s11412-009-9070-z

Sarmiento, J. W., & Stahl, G. (2008). Extending the joint problem space: Time and sequence as essential features of knowledge building. *Proceedings of the 8th International Conference on the Learning Sciences* (ICLS ’08), 24–28 June 2008, Utrecht, Netherlands (Vol. 3, pp. 295–302). International Society of the Learning Sciences. http://www.gerrystahl.net/pub/icls2008johann.pdf

Schneider, B., Worsley, M., & Martinez-Maldonado, R. (2021). Gesture and gaze: Multimodal data in dyadic interactions. In U. Cress, C. Rosé, A. F. Wise, & J. Oshima (Eds.), *International Handbook of Computer-Supported Collaborative Learning* (pp. 625–641). Springer. https://doi.org/10.1007/978-3-030-65291-3_34

Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., Frank, K., Rupp, A. A., & Mislevy, R. (2009). Epistemic network analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media*, *1*(2), 33–53. https://doi.org/10.1162/ijlm.2009.0013

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. https://doi.org/10.1111/bjet.12993

Siposova, B., & Carpenter, M. (2019). A new look at joint attention and common knowledge. *Cognition*, *189*, 260–274. https://doi.org/10.1016/j.cognition.2019.03.019

Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge*. The MIT Press. https://doi.org/10.7551/mitpress/3372.001.0001

Syczewska, E. M. (2010). Empirical power of the Kwiatkowski-Phillips-Schmidt-Shin test. In *Working Papers 45,* Department of Applied Econometrics, Warsaw School of Economics. https://ideas.repec.org/p/wse/wpaper/45.html

Tan, S. C., Wang, X., & Li, L. (2022). The development trajectory of shared epistemic agency in online collaborative learning: A study combing network analysis and sequential analysis. *Journal of Educational Computing Research*, *59*(8), 1655–1681. https://doi.org/10.1177/07356331211001562

Teasley, S. D., & Roschelle, J. (1993). Constructing a joint problem space: The computer as a tool for sharing knowledge. In S. P. Lajoie & S. J. Derry (Eds.), *Computers as Cognitive Tools* (pp. 229–258). Routledge.

Volet, S., Vauras, M., Salo, A.-E., & Khosa, D. (2017). Individual contributions in student-led collaborative learning: Insights from two analytical approaches to explain the quality of group outcome. *Learning and Individual Differences*, *53*, 79–92. https://doi.org/10.1016/j.lindif.2016.11.006

Worsley, M. (2014). Multimodal learning analytics as a tool for bridging learning theory and complex learning behaviors. *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (MLA '14), 12–16 November 2014, Istanbul, Turkey (pp. 1–4). ACM Press. https://doi.org/10.1145/2666633.2666634

Zheng, J., Xing, W., & Zhu, G. (2019). Examining sequential patterns of self- and socially shared regulation of STEM learning in a CSCL environment. *Computers & Education*, *136*, 34–48. https://doi.org/10.1016/j.compedu.2019.03.005

Zhou, G., Moulder, R., Sun, C., & D'Mello, S. (2022). Investigating temporal dynamics underlying successful collaborative problem solving behaviors with multilevel vector autoregression. In A. Mitrovic & N. Bosch (Eds.), *Proceedings of the 15th International Conference on Educational Data Mining* (EDM2022), 24–27 July 2022, Durham, UK (pp. 290–301). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.6853137

Zhou, Y., & Kang, J. (2022). Characterizing joint attention dynamics during collaborative problem-solving in an immersive astronomy simulation. In A. Mitrovic & N. Bosch (Eds.), *Proceedings of the 15th International Conference on Educational Data Mining* (EDM2022), 24–27 July 2022, Durham, UK (pp. 406–413). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.6852988