Modeling Team-level Multimodal Dynamics during Multiparty Collaboration

Lucca Eloy
Institute of Cognitive Science, University of Colorado Boulder, Boulder, CO, USA
lucca.eloy@colorado.edu

Angela E.B. Stewart
Institute of Cognitive Science, University of Colorado Boulder, Boulder, CO, USA
angela.stewart@colorado.edu

Mary J. Amon
Institute of Cognitive Science, University of Colorado Boulder, Boulder, CO, USA
mary.amon@colorado.edu

Caroline Reindhardt
Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, CO, USA
caroline.reinhardt@colorado.edu

Amanda Michaels
School of Social and Behavioral Sciences, Arizona State University, Glendale, AZ, USA
apmichae@asu.edu

Chen Sun
Department of Educational Psychology and Learning Systems, Florida State University, Tallahassee, FL, USA
cs15c@my.fsu.edu

Valerie Shute
Department of Educational Psychology and Learning Systems, Florida State University, Tallahassee, FL, USA
vshute@fsu.edu

Nicholas D. Duran
School of Social and Behavioral Sciences, Arizona State University, Glendale, AZ, USA
nduran4@asu.edu

Sidney K. D'Mello
Institute of Cognitive Science, University of Colorado Boulder, Boulder, CO, USA,
sidney.dmello@colorado.edu

ABSTRACT

We adopt a multimodal approach to investigating team interactions in the context of remote collaborative problem solving (CPS). Our goal is to understand multimodal patterns that emerge and their relation with collaborative outcomes. We measured speech rate, body movement, and galvanic skin response from 101 triads (303 participants) who used video conferencing software to collaboratively solve challenging levels in an educational physics game. We use multi-dimensional recurrence quantification analysis (MdRQA) to quantify patterns of team-level regularity, or repeated patterns of activity in these three modalities. We found that teams exhibit significant regularity above chance baselines. Regularity was unaffected by task factors. but had a quadratic relationship with session time in that it initially increased but then decreased as the session progressed. Importantly, teams that produce more varied behavioral patterns (irregularity) reported higher

emotional valence and performed better on a subset of the problem solving tasks. Regularity did not predict arousal or subjective perceptions of the collaboration. We discuss implications of our findings for the design of systems that aim to improve collaborative outcomes by monitoring the ongoing collaboration and intervening accordingly.

CCS CONCEPTS

• **Human-centered computing~Empirical studies in collaborative and social computing**

KEYWORDS

Multimodal interaction, Behavioral Regularity, Collaborative Problem Solving, MdRQA

ACM Reference Format:

Lucca Eloy, Angela E.B. Stewart, Mary J. Amon, Caroline Reinhardt, Amanda Michaels, Chen Sun, Valerie Shute, Nicholas D. Duran, and Sidney K. D'Mello. 2019. Modeling Team-level Multimodal Dynamics during Multiparty Collaboration. In *Proceedings of the 21st ACM International Conference on Multimodal Interaction (ICMI '19), Oct. 14-18, 2019*, Suzhou, China. https://doi.org/10.1145/3340555.3353748

1 Introduction

Imagine you are a software developer. You wake up and coordinate with your partner to get the kids to school on time without being late for work. At the office, you work with your team of developers to fix bugs in a new product, and then collectively decide what to get for lunch. After lunch, you have a Skype meeting with the marketing department to discuss the product launch campaign. You head to the park after work for a game of basketball with your friends before getting home and planning out the family's schedule for the rest of the week. Throughout the day you spontaneously collaborate with various people to get things done.

Be it in school, work, family, or other social groups, we are constantly engaging in interaction and coordination with people around us to accomplish a variety of tasks involving others [68, 69]. Indeed, human-human interaction is the vehicle by which teams work to achieve a shared goal that may not be possible by working alone [16, 23]. A marketing team working on an ad concept and a group of friends playing basketball both require individuals to contribute their knowledge and skills to achieve the goals of the group through meaningful interaction.

Although multiparty interaction is ubiquitous in everyday life, people often fail to navigate such interactions efficiently. Successful communication relies on socio-cognitive-affective processes such as conversational grounding, turn-taking, emotion co-regulation, and joint attention, to name a few [8, 27, 39, 48, 51, 59]. Disruption of these processes often occurs, resulting in loss of coordination within a group [31, 63]. Additional factors that undermine a group's effectiveness, including social loafing, where people tend to put less effort into a task when working in groups, and group think, where groups make suboptimal decisions due to diminished individual responsibility or social pressures [4, 57]. There have been efforts to support multiparty interaction, for example, by increasing individual accountability [67, 90]. However, there is little research on how productive and unproductive group processes can be computationally assessed, interpreted, and intervened upon. We address gap by uncovering the multimodal dynamics of multiparty interaction in the context of collaborative problem solving (CPS) and understanding how these dynamics predict outcomes.

1.1 Multimodal signals in collaborative problem solving

Problem solving is defined as cognitive processing required to understand and resolve a problem when the solution is not obvious [69]. Collaborative problem solving occurs when two or more people engage in a coordinated attempt to share their skills and knowledge in order to construct and maintain a joint solution to a problem [74]. Effective CPS is dependent on a team's ability to establish common ground [74], jointly develop a solution that accommodates multiple perspectives [74], and monitor progress toward the goal [29, 75]. Given that CPS skills are essential in the 21st century globalized workforce, researchers [15, 68, 69, 76] are addressing this need by emphasizing CPS in educational curricula [12, 19, 36, 38, 43, 68, 69]. For

example, the Program for International Assessment (PISA) made CPS assessment one of their focal areas in 2015, where students from over 70 countries were assessed on their CPS skills [69].

Despite this emphasis on assessing CPS skills, we know very little about how to promote or support effective CPS. This is a major challenge since teams often perform worse than they potentially could, a phenomenon known as "process loss" [22, 23, 72]. Process loss is exacerbated during remote interactions as basic social signals like speech, breathing, gesturing, and other subtle movements are dampened due to limitations with camera resolution, bandwidth, audio feedback, and delayed sound transmissions [80]. Degradation in these lower-level cues reduces a team member's ability to perceive, integrate, and extrapolate meaning from the interaction, resulting in lower coordination, cohesion, trust, and performance among team members [1, 80]. Thus, it is crucial to study the behavioral dynamics of remote CPS interactions to better understand factors that contribute to team performance. This understanding can then help develop computerized supports to facilitate collaboration and its outcomes.

Importantly, a multimodal approach might be necessary because different modalities index different aspects of the group interaction. High-level language (e.g., specific suggestions) communicate the content of the collaboration [29], while lower-level language products like speech rate index turn-taking dynamics [34]. Social cues and emotional aspects of the collaboration are conveyed through acoustic-prosodic information [28] (e.g., pitch) and facial expressions [26]. Head movements have unique signaling capacities in communication as they are highly visible and involved in conversational feedback (e.g., acknowledgment uptakes) and used to underscore semantic content [47]. Physiological signals (e.g., heart rate, galvanic skin response) index internal states like autonomic arousal [25], and have been linked to emotion co-regulation [40, 50] and improved performance in simple problem solving tasks [32, 41, 42].

1.2 A multimodal dynamical systems approach to understanding CPS

We adopt a multimodal approach to model the various modalities involved in CPS [33, 66, 94]. We ground our approach in nonlinear dynamical systems theory, which emphasizes interactive relationships that result in nonlinear shifts in behavior over time [24, 58]. Such patterns are typical to CPS, where teams exhibit qualitatively distinct behaviors, for example, as they move from discussing possible strategies to solving a problem, to executing a select strategy, to evaluating its effects. In line with this perspective, we utilize an analytical method from nonlinear dynamical systems theory called recurrence quantification analysis (RQA) [70] to quantify behavioral patterns which emerge during the collaboration [18].

The basic idea of RQA is to identify the extent to which a dynamical system exhibits recurrent or repeated states of activity, which can provide critical insights into the structural and temporal organization of the system. RQA captures points during which a state is revisited, as well as sequences of states that are repeated. It can be used to examine recurrence within one signal or among multiple signals, where it measures coordination [18, 70, 83]. RQA has many advantages over traditional time series analysis techniques because it can be applied to signals of any length and has no assumptions of linearity or stationarity [84].

RQA has traditionally been used to quantify recurrent patterns within one signal (e.g., [89, 93]) or to study alignment between two signals (cross-RQA (e.g., [18, 40, 46, 50, 71, 73, 83]). Here, we use multidimensional RQA (MdRQA) [92], a newer extension of RQA, to examine patterns across multiple signals from multiple people and modalities. MdRQA quantifies the degree to which the collective organization of the signals exhibits *regular* patterns of behavior, which reflect periods when the system is in the same repeated "state," though the individual channels may not be in alignment [3]. For example, one repeated state might include high physiological arousal but low speech rate and negligible bodily movement (potentially signaling tension), whereas another might include high physiological arousal with rapid speech and frantic movement (potentially signaling excitement). Here, we use MdRQA to examine the multimodal dynamics of multiparty interaction amongst triads engaged in a remote CPS task.

1.3 Related work

We review literature pertaining to models of effective collaboration, interpersonal synchrony and coordination, and recent studies that have used MdRQA to investigate team dynamics. With respect to the first, prior work has investigated how team composition (e.g., group size [55, 56], ability/motivation cohesiveness [27, 37, 52]) and features of the task (e.g., problem structure [54, 81]) predict team performance. Taking a social psychology approach, these studies primarily focus on group and task features that exist before the interaction rather than the dynamics of the interaction itself. Subsequently, computational approaches have been applied to predict stable individual traits related to teamwork, such as personality [4], leadership and dominance [11, 45, 78], and interpersonal skills (empathy, communication) [44, 65]. Again, the focus is on predicting attributes of individuals from their behavioral patterns.

In the HCI domain, research has prototyped systems that deliver automated feedback based on interaction performance [49, 77, 88]. For example, following a group task in a video conferencing environment, the Collaboration Coach (CoCo) [77] provides participants with information on their own turn-taking, affect, and participation during the task. This type of feedback has been shown to improve collaborative outcomes, such as task performance [88] or individual participation [49, 77].

More recent research has shifted to examining dynamic processes underlying collaborations. In this domain, studies on multiparty interaction have attempted to predict group performance using machine learning. For example, Avci and Aran [6] used modified Hidden Markov Models [9] to predict group performance from speech cues (turn-taking, interruptions, etc.) and head and body movement features. Murray & Oertel [62] followed up on this work using a transfer learning approach with Random Forest classifiers. They extracted acoustic features from audio recordings and linguistic features from audio, to predict task performance with a mean-squared error of 64.4 (baseline = 79.3).

Rather than predicting phenomena related to group interaction (e.g., task performance [6, 62] or empathy [44]), the extensive literature on coordination aims to describe patterns of the interaction itself. This work quantifies the alignment of signals between interacting individuals [20, 33, 66], with an emphasis on verbal [94], physiological [66], and behavioral (e.g., movement [79] and eye gaze [72]) signals. For example, Von Zimmerman and Richardson [94] showed how verbal coordination can strengthen large group affiliation and performance. Similarly, motor coordination has been shown to positively influence perception of affiliation [53]. Eye gaze coordination (a proxy for joint attention) has been shown to increase when common ground is established prior to the conversation and is linked to better outcomes [72].

In a somewhat different vein, some research has developed computational models to predict individual behaviors of one team member from behaviors of others in the team, arguing that this provides a more direct and generalizable measure of coordination compared to traditional analytical approaches. In particular, Grafsgaard et al. [35] trained long-short term memory (LSTM) networks to predict facial expressions and body and gestural movement in heterosexual romantic couples, using behavioral inputs from one partner to predict those of the other. Applying a similar approach to triadic CPS interaction, Stewart et al. [87] trained LSTMs to predict speech rate of one teammate from speech rate and acoustic-prosodic features of two others and from team-level task context features. They could predict speech rate up to six seconds in advance, arguing that the modes are able to capture co-regulation in turn-taking dynamics.

Most recently, researchers have looked beyond coordination patterns, to quantify regularity, or recurrent patterns of activity across team members [64, 89, 93]. More closely related to our work, Vrzakova et al. [91] investigated patterns of team-level regularity in triads engaging in a remote, collaborative programming task. They used MdRQA to quantify recurrent (repeated) patterns amongst three individuals' eye-gaze in tandem with changes on a shared user interface. They found that regularity (i.e., recurrence) in these four signals positively predicted expert-codes of team negotiation and coordination, which is defined as the process of collaboratively settling on a solution and executing it.

Similarly, Amon et al. [3] applied MdRQA on an expanded version of the same data set to examine regularity in eight channels: speech rate for three individuals, body movement for three individuals, and activity in two areas of interest on a shared user interface. They found that regularity within short time periods (i.e., one to two seconds) was significantly higher than shuffled baselines but did not systematically change over the course of the 20-minute collaboration. Most importantly, they found an inverse relationship between regularity and expert codes of negotiation and coordination and construction of shared knowledge (the process of expressing ideas and understanding other's ideas). That is to say, novel behaviors (lower regularity), predicted behaviors associated with successful collaborations.

The contradictory findings in these two studies is attributable to differences in modality. Whereas regularity in Vrzakova et al. [91] essentially indexed joint attention in tandem with a shared display, a desirable process, regularity in the Amon et al. [3] study was related to an overall reduction in behaviors, which is ostensibly undesirable. Neither of these previous studies were able to reliably link regularity with CPS outcomes.

1.4 Current study, research questions, and novelty

Research on multiparty interaction, including the CPS literature, has typically focused on modeling traits of the interacting individuals (e.g., personality [2, 6]), processes involved in the interaction (e.g., turn-taking [44]), or the outcomes of the interaction (e.g., task performance [6, 62]). Our work is novel in that it instead focuses on multimodal interaction patterns emergent in team interactions, identifying factors that might influence these patterns, and investigating the extent to which the patterns index collaborative outcomes. Further, we focus on analyzing the team as a whole, which is a departure from previous research that focuses on individual team members or alignment between team members. Although automated feedback systems are a potential application of our work, the present study focuses on the more preliminary step of understanding team-level multimodal dynamics.

We collected a large data set of 101 triads (303 participants) who used video conferencing software to collaboratively solve challenging levels in an educational physics game. We then use multidimensional recurrence quantification analysis [92] (MdRQA) to jointly model three team-level signals: speech rate to index verbal contribution, body movement features to index nonverbal communication, and galvanic skin response as a measure of autonomic arousal. We selected these three modalities as they have been linked to coordination amongst interacting individuals [35, 66, 94] and are known to index key CPS processes including active participation [48], turn-taking dynamics [48], shared attention [5], and emotional co-regulation [17]. Our key measure is multimodal regularity which reflects the extent to which the team's behavior/physiology exhibits recurrent (or repeat) versus irregular (potentially novel) behaviors, which is a measure derived from dynamical systems theory (see Section 1.2).

We address five research questions. First, *do teams exhibit systematic patterns of multimodal regularity* (RQ1)*?* A positive finding here would reflect an underlying systematicity in teams' behavioral/physiological patterns, providing a basis for our remaining questions. Next, we examine how overall activity in individual modalities corresponds to multimodal regularity by asking *how do individual measures constitute team-level regularity* (RQ2)*?* For example, does higher recurrence correspond to increased speech rate, less body movement, and elevated galvanic skin response (GSR) signals, or some other organization of these component modalities? Next, research has found that changes in context affect multiparty interactions [54, 81] so we investigate *how regularity varies as a function of task features* (RQ3) by examining whether it is robust or malleable to differences in the task context. Further, we examine temporal patterns of regularity by addressing *how does regularity change over time?* (RQ4). It could be the case that teams establish stronger patterns of interaction as the collaboration progresses or, alternatively, time could play a negligible role in interactions as patterns could rapidly form and stabilize. Finally, we address the question of *whether regularity predicts subjective and objective team outcomes net of overall levels of behavior* (RQ5).

Although using MdRQA in the context of multiparty interaction is relatively novel, this approach has been used in two studies discussed above [3, 91]. We distinguish our research from these in the following four ways. First, we incorporate galvanic skin response, a physiology signal that indexes autonomic arousal [21,

66] and has been linked to emotional arousal [25, 61]. Measures of peripheral physiology have been extensively studied in the interpersonal coordination literature [66] and provide additional insight into the internal state of interacting individuals beyond outward behavioral measures used in these previous studies (e.g., eye gaze, speech rate, body movement). Second, previous work examined a single task context with one goal and a single team interaction. In our work, we manipulate the goals and context of the task and analyze two successive collaboration sessions with an eye towards quantifying the extent to which dynamics vary as a function of task context changes. Third, previous studies analyzed interactions of signals from individuals and not the team as a whole, thus they might not fully capture team-level dynamics as in the current study where we aggregate signals from individual team members to obtain a team-level representation. Finally, previous studies were insufficiently (statistically) powered (19 and 32 teams were analyzed) and found no significant link between regularity and task performance. With our larger dataset of 170 collaborative sessions across 86 teams, we for the first time, demonstrate the relationship between regularity and CPS outcomes. In short, the present study examines how multimodal, team-level processes, both those internal and external to team members, interact to produce systematic behavior, and the extent to which this corresponds to team outcomes.

2 Data Collection

2.1 Participants

Participants were 303 students (56% female, average age = 22 years) from two large public universities. Students were 47% Caucasian, 28% Hispanic/Latino, 18% Asian, 2% Black or African American, 1% American Indian or Alaska Native, and 4% other. Students were assigned to 101 teams of three based on scheduling constraints. Thirty students from 18 teams (26%) indicated they knew at least one person from their team prior to participation. Participants were compensated monetarily ($50) or with course credit after completing both the at-home and in-lab portions of the study.

2.2 Problem solving environment

We used Physics Playground for our problem solving environment. This is a two dimensional educational game that aims to teach students basic Newtonian physics concepts (e.g., Newton's laws, energy transfer, and torque) and has been found to be highly engaging [13, 60]. Everything in the game obeys the laws of physics. Game levels require participants to guide a green ball to a red balloon by drawing simple machines (i.e., ramps, levers, pendulums, and springboards) using the mouse. For example, Figure 1 depicts a team's solution that involved drawing a lever that rotated around a fixed point and weighting down one end in order to roll the ball toward the balloon. Teams earned gold trophies for more concise solutions where fewer objects were drawn; otherwise, their solution earned a silver trophy. Players could restart, exit, or change levels at any time during gameplay. However, there were no hints or support mechanisms in the game with the exception of a tutorial on game mechanics. Each game level was rated by experts based on physics knowledge required to solve the level and difficulty of game mechanics.

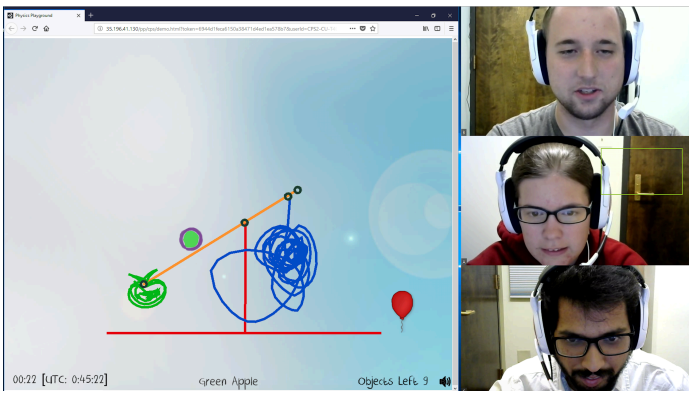

Figure 1: Screenshot of the task setup, showing a triad attempting to use a lever and weight to navigate the green ball to the red balloon.

2.3 Procedure

All procedures were approved by the institution's Institutional Review Boards. Participants completed informed consent forms prior to the lab portion of the study.

2.3.1 At-home procedure

Participants were emailed a survey to complete prior to the in-lab portion of the study. The survey was sent to them at least 24 hours in advance of their scheduled lab session. The online survey included demographic questionnaires assessing participants' gender, age, and prior physics experience. Participants also completed a ten-item expert-created physics pre-test that assessed knowledge of energy transfer and properties of torque, which corresponded to the Physics Playground levels selected for the in-lab portion of the study.

After completing the physics pre-test, participants were provided a short tutorial describing how to use the Physics Playground environment, including how to draw machines such as ramps and springboards. Participants then had up to 15 minutes to complete five expert-rated easy levels to familiarize themselves with the game. Other activities in the in-home survey not germane to the present study are not discussed.

2.3.2 In-lab procedure

Upon arriving to the lab, participants were each assigned to one of three separate computer-enabled workstations either partitioned or in different rooms (depending on the school) with video conferencing capabilities and screen sharing through Zoom (https://zoom.us) (Figure 1). Each computer had a webcam with a headset microphone so students could see and hear each other. Audio was also recorded from this headset, and individual participants were video recorded using a second webcam. Physiological responses were recorded using the Shimmer 3 GSR+ to measure GSR at 51.2 Hz. The experimenter attached the GSR electrodes to the palm of each participant's non-dominant hand and attached the receiver device on their non-dominant arm. An ear clip that measures photoplethysmography (PPG) was attached to the ear lobe on the same side of the body. These electrode placements followed the manufacturer's recommendations.

Teams completed the Physics Playground task in three 15-minute blocks, including a warmup and two experimental blocks. Only a single randomly assigned team member could interact with the game, and this participant's screen was shared using zoom's screen sharing mechanism. A different team member was given control of the mouse and interface during each 15-minute block such that each participant controlled the interaction for exactly one block.

Participants first completed the 15-minute warmup as a team in order to familiarize them with their team members and the task environment. They were instructed verbally and with on-screen instructions to use the 15 minutes to get to know their teammates and to play a few levels together. They were given five easy-to-medium levels corresponding to the energy transfer and torque physics concepts. Teams were given on-screen warnings when they had ten and five minutes left in each 15-minute block.

After the warmup, screen sharing was disabled and participants were *individually* asked to rate their emotional valence (1 = *very negative*, 5 = *very positive*) and arousal (1 = *very sleepy*, 5 = *very active*) on five-point Likert scales. They also completed a six-item questionnaire assessing perceived quality of their team's collaborative problem solving (e.g., "Our team freely shared our thoughts about the nature of the problem and possible solutions"). This was followed a three-item inclusiveness and team norms questionnaire (e.g., "Each person on my team had an equal say in the decisions made during gameplay") (1 = *disagree strongly*, 7 = *agree strongly*) [30].

Next, teams collaborated during two experimental blocks with differing goals. In one goal manipulation, teams were instructed to "solve as many levels as possible." In the other goal manipulation, teams were instructed to "get as many gold trophies as possible." Teams were reminded that they earned a gold trophy by using fewer objects in their solution. Instructions with the goal condition were provided both verbally and on screen. The purpose of this manipulation was to examine the degree to which task constraints influenced teams' patterns of behavior. The physics concept was also manipulated during the experimental blocks with teams being presented with either seven energy transfer levels or six properties of torque levels to complete. All levels were of medium-to-hard difficulty. Goal manipulation and physics concept were counterbalanced across teams. For example, a team that started with energy transfer levels with the goal of obtaining as many gold trophies as possible in the first experimental block would then be instructed to solve as many levels as possible in the second block while being presented with torque levels. Another team would start with the energy transfer concept while tasked with solving as many levels as possible in the first experimental block, followed by the torque concept and trophies goal in the second experimental block. Teams were given 15 minutes for each experimental block and received the same on-screen warnings as the warmup when they had ten and five minutes left in the 15-minute block. They were also reminded of their goal condition (levels or golds) during this warning. After each experimental block, participants *individually* (i.e., without screen sharing) completed the same surveys that they completed after the warmup.

Additionally, after both experimental blocks, participants individually completed a physics post-test, which was a parallel-form version of the pre-test. Assignment of test version (A or B) as pre- or post-test was counterbalanced across participants. Lastly, participants engaged in an unrelated collaborative task, which is excluded from the current analyses. They were then fully debriefed.

2.3.3 Manipulation check

To ensure that the goal manipulations were sufficient, we asked each participant individually after each experimental block to report their primary goal for that block. Overall, 95% of responses were consistent with the intended manipulation, indicating that our manipulation was successfully perceived by the vast majority of teams.

3 Data Processing

3.1 Deriving time series

Given that the warmup was focused on familiarizing teams with their team members and the task environment, we focused our analyses only on the two experimental blocks. First, we used the IBM Watson Speech to Text service [36] to generate transcriptions of individual audio recordings, from which we computed speech rate (words per second) for each second of the collaboration, yielding a 1Hz time series. If a word spanned multiple seconds, we assigned it to the second in which it started.

Videos of the participants' faces and upper bodies were recorded at a variable frame rate. We converted them to a constant frame rate of 10 fps using FFmpeg and computed face and upper movement from these videos using a validated motion estimation algorithm [15]. The algorithm computes the proportion of pixels in each frame that change from a continuously updated background image from the previous four frames. Estimates of face and upper movement were smoothed by averaging over five consecutive frames resulting in a 2 Hz time series.

GSR was recorded at 51.2 Hz., which is a relatively low frequency signal compared to some research-grade systems. Therefore, we focused our analyses on the slower-moving tonic components of the signal as this is preserved at lower sampling rates [14]. To extract this, GSR series were first smoothed using a cubic smoothing spline (smoothing parameter = 0.25) [86] followed by a second-order low-pass Butterworth filter with a cutoff frequency of 5 Hz to remove motion artifacts [82]. This is consistent with literature suggesting the tonic signal changes at a rate slower than 5 Hz [14, 70, 82].

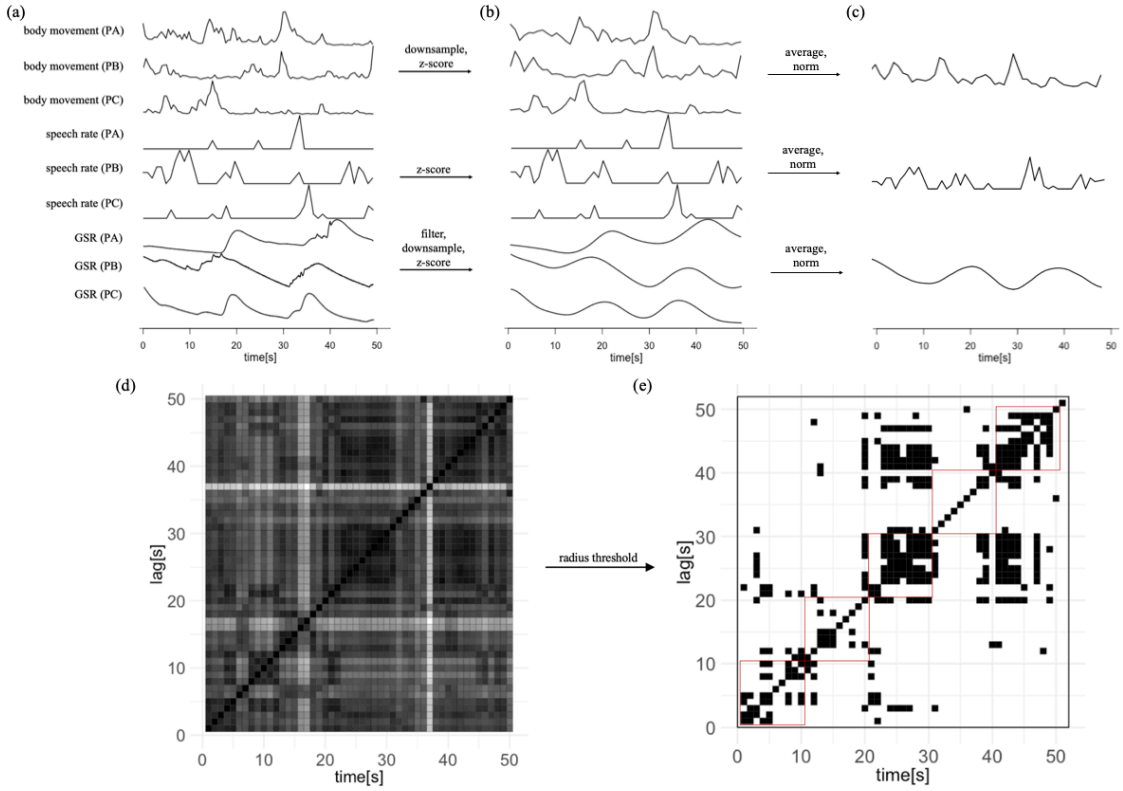

Figure 2: (a) Example time series for three participants undergoing (b) pre-processing and (c) team-level aggregation before the MdRQA. (d) Distance matrix computed from team-level time series transformed into (e) binarized recurrence matrix obtained after applying a radius, with black points representing recurrent points. Red boxes illustrate how windowed (10 second) recurrence rates are obtained.

3.2 Downsampling, standardization, and aggregation

Our analyses required time series with equivalent sampling rates. Turns, defined as spoken utterances, were quite short (median of 1.4 seconds), so we chose a 1 Hz sampling rate for modeling to capture team dynamics at fine-grained intervals. Speech rate was initially measured at a 1 Hz frequency, thus did not require resampling. Body movement and GSR were down-sampled to 1 Hz with an order eight Chebyshev type I filter using the R function 'decimate' in the 'signal' package [85]. To account for individual differences, speech rate, body movement, and GSR signals were *z*-score standardized for each participant. We then averaged the time series across participants in the team, resulting in one time series per modality per team, to analyze the team as a single unit. To account for differences in value ranges across modalities, we normalized the team-level time series to a range of 0 to 1. Figures 2a, 2b, and 2c illustrate the steps of data pre-processing per modality.

3.3 Data exclusion

Occasionally the GSR signal would have missing data due to the lightweight sensor which could lose contact with the skin from too much movement. Some data loss is expected since movement was not restricted to

preserve ecological validity of the task. Poor audio and video recordings also resulted in missing speech rate and body movement data. To account for this, we excluded participant signals that did not have a complete 15 minutes of data at a constant sampling rate (i.e., chunks in the time series were missing). If more than one participant was missing data in a single modality, the entire team was excluded from analysis. For example, if participant A in a team had missing audio and video recording, but participants B and C had intact data, participant A's speech rate and body movement data would be excluded from analysis. If participant B also had missing audio recording, the entire team would be excluded. We chose this exclusion criteria to maintain a large sample size without sacrificing data integrity as teams with excluded data retained at least 6 out of 9 valid participant-level time series. Overall, 83% of analyzed teams had complete data. To control for this difference, we grouped teams based on data completeness and z-scored recurrence rates per group.

Teams that did not complete the at-home or in-lab surveys were excluded due to missing outcome measures. Of the original 101 teams, 14 were entirely excluded for missing data and two teams had one experimental block excluded for missing data. This resulted in a final analytic sample of 258 participants, or 86 teams, who completed a total of 170 experimental blocks.

3.4 Measures

3.4.1 Outcome measures

The individual-level perceived collaboration quality and team norm questionnaires [30] had a Cronbach's alpha reliability of 0.84 and 0.91, respectively. Combining the scales corresponded to an alpha reliability of 0.93. Thus, we averaged item scores across the two questionnaires to obtain a single composite score of perceived collaboration quality per individual. We then averaged the composite score across team members, resulting in one aggregate score per team. Valence and arousal were also averaged across team members.

We also computed a task performance score for each block by dividing the total number of trophies a team earned by the number of possible trophies (i.e., seven for energy transfer or six for torque). We simply added gold and silver trophies and did not distinguish between the two, as trophy type is related to the goal manipulation and not central to the present study. We also did not include post-test scores because they cannot be easily resolved at the team level as they are highly dependent on pre-test scores.

3.4.2 Multimodal Recurrence Quantification Analysis (MdRQA) measures

Multi-dimensional recurrence quantification analysis (MdRQA) [92] takes a multi-column matrix as input, with each column representing a single time series. MdRQA computes a distance matrix by calculating pairwise Euclidean distances between the multidimensional values at each time point. The diagonal of the distance matrix represents the line of identity (LOI), or the distance between each time point with itself (i.e., lag 0). Diagonals of the distance matrix parallel to the LOI represent distances between elements at various time lags, with diagonals further from the LOI occurring at greater lags.

A radius is applied to transform the distance matrix into a binary recurrence matrix. Elements in the matrix whose value falls below the radius are considered recurrent points and assigned a value of one. Distances that are equal to or greater than the radius are assigned a value of zero, denoting no recurrence. We estimated the radius parameter with a grid search on a randomly-sampled 25% of teams (both experimental blocks) to obtain a target average recurrence rate in the recommended 4% and 5% range [18]. This yielded a radius of 0.26.

Recurrence rate is the percentage of recurrent points in the matrix and is taken as our primary measure of team-level regularity. In our case, a recurrent point means that the team-level values of speech rate, body movement, and GSR at a given second return to a repeat state. In addition to identifying total recurrence across the task session, we also studied changes in recurrence over time within an experimental block by computing recurrence rate per minute or each 60-second square window along the LOI. Figures 2d and 2e illustrate the steps of calculating an MdRQA recurrence matrix and obtaining windowed recurrence rates. We did not consider additional measures that can be derived from the recurrence plot such as determinism and entropy because they were strongly correlated (*r*s between .51 and .90) with recurrence rate.

3.4.3 Overall activity measures

We also computed *average team activity* for each modality. This was calculated by summing each participant's time series per modality and then averaging across the three participants in the team. For speech rate, this translates to the average of the total words spoken by participants, or average team verbosity. Body movement and GSR were also summed within each participant and then averaged across team members. To account for differences in the range of each modality, we *z*-scored each measure across all teams.

4 Results

Figure 3a shows a histogram of recurrence rates (RR) pooled across both blocks. Figure 3b, 3c, and 3d depict sample recurrence plots corresponding to low (RR = 2.38%), average (RR = 4.65%), and high (RR = 9.80%) recurrence. Table 1 presents descriptive statistics of all key study variables.

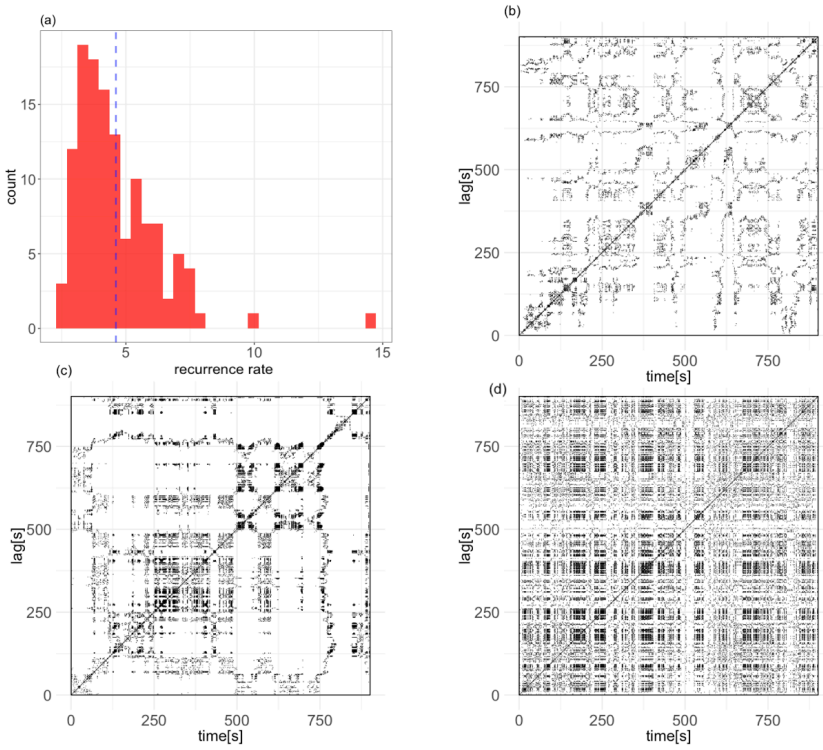

Figure 3: (a) Histogram of observed recurrence rates before z-scoring. Blue line indicates the mean recurrence rate across teams. (b) Recurrence plots for low RR team (RR = 2.38%), (c) average RR team (RR = 4.65%), and high RR team (RR = 9.80%). Black points denote recurrent points.

Table 1: Descriptives of key team-level variables.

Measures	Mean (SD)	Observed Range	Possible Range
MdRQA Recurrence Rate (%)	4.70 (1.63)	[2.51 - 14.81]	[0 - 100]
Perceived Collaboration Quality	6.40 (0.45)	[4.52 - 7.00]	[1 - 7]
Perceived Valence	3.85 (0.75)	[1.33 - 5.00]	[1 - 5]
Perceived Arousal	3.52 (0.68)	[1.67 - 4.67]	[1 - 5]
Proportion of Trophies Earned	0.22 (0.19)	[0.00 – 0.75]	[0 - 1]

4.1 (RQ1) Do teams exhibit multimodal regularity?

Our first research question pertained to whether teams exhibited systematic patterns in regularity. To address this, we compared observed MdRQA recurrence rates with shuffled baselines, obtained by randomly shuffling the 1s time points of each team's multi-dimensional time-series such that concurrent values across channels remained together but the ordering of time points was randomized. Shuffled time series were then submitted to a MdRQA (with the same radius as the observed data) to derive a measure of baseline regularity. We computed recurrence rate for ten shuffled baseline time series and averaged across them.

Since MdRQA computes pairwise distance between all points, shuffling does not affect the overall recurrence rate. However, local dependencies should be disrupted in the shuffled time series. To test this, we computed the proportion of recurrence points for both the original and the shuffled time series at consecutive lines parallel to the main diagonal, which represent successive lags of 1s, 2s, etc. We investigated lags of up to 10s to capture a range of possible local patterns and ignored lag 0s since it involves comparing each point to itself, which will always be recurrent. For systematicity to exist, observed recurrence rates should differ from baselines, suggesting that there is significant multimodal regularity (higher recurrence rates) or irregularity (lower recurrence rates).

Figure 4a depicts average observed recurrence rates across teams at individual lags against the shuffled baselines. Two-tailed paired-sample *t*-tests comparing observed to baseline team recurrence rates (averaged over the two experimental blocks) indicated that the observed recurrence rates at lags 1 - 10 were significantly higher than baselines ($ps < 0.001$ with a Bonferroni correction resulting in an effective significance threshold of 0.005 [.05/10]). Regularity was highest at lag 1, upon which it decreased but was still significantly higher than the baselines at all lags. This suggests that teams exhibit higher regularity in shorter intervals, but patterns remained at further time lags.

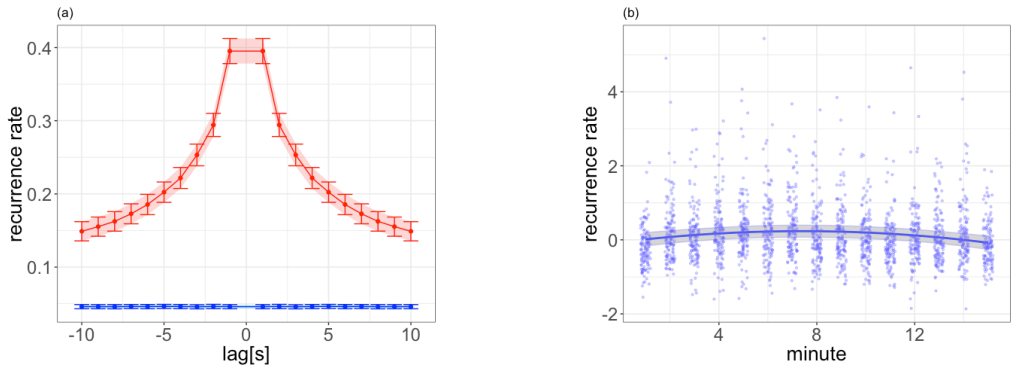

Figure 4: (a) Observed MdRQA recurrence rates at time lags 1-10s (red) versus randomly shuffled baselines (blue). Error bars indicate 95% confidence intervals. (b) Quadratic trend of time (minute) on regressing recurrence rate. Band around the fitted line denotes confidence interval, and blue points represent observations.

4.2 (RQ2) How do individual measures constitute regularity?

We used a linear mixed-effects modeling approach with each team as a random intercept for all subsequent analyses. This is the recommended approach due to the repeated (multiple blocks) and nested (blocks nested in teams) nature of the data. We used the 'lme4' library in R [10] for the requisite computation.

To identify the relationship between individual measures of teams' behavior and team-level regularity, we regressed MdRQA recurrence rate on average team-level measures of speech rate, body movement, and GSR. Goal manipulation (solve levels or earn trophies), physics concept (torque or energy transfer), experimental block number (first or second), and teams' school affiliation (two groups) were included as covariates to control for contextual factors. We also included goal manipulation (two groups) and physics concept (two groups) as task-level covariates (see below). Results indicated that higher recurrence rate was associated with

lower speech rate (B = -0.53, p < .01), GSR (B = -0.24, p < .01), and body movement (B = -0.20, p < .01). Thus, reduced activity in each modality corresponded to more repeat values and higher recurrence rate. Conversely, lower regularity, or irregularity, corresponded to increased activity in the component behavioral time series.

4.3 (RQ3) How do features of the task affect regularity?

We investigated the extent to which features of the task predicted team-level regularity. The model from Section 4.2 also tested the effects of goal manipulation (B = -0.01, p = 0.90) and physics concept (B = -0.16, p = 0.10). Neither of these variables showed a significant effect on recurrence rate, indicating that team regularity was stable with respect to task constraints.

4.4 (RQ4) How does regularity change over time?

To examine the degree to which teams' regularity varied over time, we regressed minute-level recurrence rate, computed as the percent of recurrent points in each minute of the interaction (Figure 2e) on time (minute), testing for both linear and quadratic trends. We included average team movement, verbosity, GSR activity, goal manipulation, physics concept, experimental block number, and school affiliation as covariates to control for these effects. We found evidence for a significant linear increase in recurrence rate (B = 0.08, p < 0.01) throughout the block. However, there was also a significant quadratic trend (B = -0.01, p < 0.01) in that participants first increased and then decreased recurrence rate as the session progressed (Figure 4b).

4.5 (RQ5) Regularity as a predictor of team performance

Lastly, we examined the degree to which regularity predicted teams' subjective and objective outcomes. For this, we separately regressed team valence, arousal, perceived collaboration quality, and proportion of trophies earned on MdRQA recurrence rate. Teams' school, experimental block, goal manipulation, and physics concept were included as covariates to account for context effects. Average team verbosity, average team movement, and average team GSR were also included as covariates to control for overall amount of activity of each team. The results are presented in Table 2.

We found that that team-level regularity negatively predicted valence (B = -0.15, p = 0.05), but not arousal (B = -0.01, p = 0.93) or perceived collaboration quality (B = -0.01, p = 0.88). For objective outcomes, we found that recurrence rate was not associated with proportion of trophies earned in the block (B = -0.01, p = 0.37). However, we did find a significant effect of physics concept on the trophies earned (B = .25, p < .01), suggesting that teams earned more trophies for the concept of torque compared to energy transfer. To explore this further, we included the regularity × concept interaction term in the model, which was significant (*ChiSq (1)* = 4.24, p = 0.04). A subsequent simple slopes analysis identified a significant negative effect of recurrence rate on proportion of trophies earned (B = -0.04, p < 0.05) for the torque concept, but not for the energy transfer concept (p = 0.84). Taken together, the findings suggest that lower levels of team regularity, or heightened *irregularity,* corresponded to more positive valence and better task scores for the easier concept.

Table 2: Results of linear mixed-effects model regressing outcome measures on MdRQA recurrence rate and several covariates. Estimates (*B*) and standard errors (*SE*) shown.

	Arousal		**Valence**		**Subjective quality of collaboration**		**Proportion of trophies earned**	
	B (SE)	*p*	*B (SE)*	*p*	*B (SE)*	*P*	*B (SE)*	*p*
MdRQA recurrence rate	-0.01 (0.07)	0.93	-0.15 (0.08)	0.05	-0.01 (0.04)	0.88	-0.01 (0.02)	0.37
Covariates								
Average team verbosity	0.07 (0.07)	0.28	-0.02 (0.07)	0.81	0.18 (0.04)	< 0.01	-0.01 (0.01)	0.64
Average team movement	-0.03 (0.06)	0.67	-0.08 (0.07)	0.20	0.02 (0.04)	0.64	-0.02 (0.01)	0.16

Average team GSR	0.12 (0.06)	0.06	0.07 (0.06)	0.27	-0.02 (0.04)	0.56	0.02 (0.01)	0.22
Goal manipulation (levels)	0.00 (0.07)	0.98	-0.01 (0.09)	0.94	-0.02 (0.04)	0.57	-0.03 (0.02)	0.21
Physics concept (torque)	0.32 (0.07)	< 0.01	0.64 (0.09)	< 0.01	0.14 (0.04)	< 0.01	0.25 (0.02)	< 0.01
Experimental block (block 2)	-0.04 (0.07)	0.58	0.15 (0.09)	0.10	0.05 (0.04)	0.29	0.03 (0.02)	0.09
School (CU Boulder)	-0.16 (0.14)	0.25	-0.08 (0.13)	0.53	0.00 (0.09)	0.96	0.02 (0.03)	0.38

5 Discussion

We investigated multimodal dynamics during remote collaborative problem solving (CPS) interactions to better understand factors of the interaction that contribute to CPS processes and outcomes. Our aim is to understand how collaboration can be improved by uncovering the dynamics of multiparty interaction and to leverage these insights to develop better technologies to support remote collaborations.

5.1 Findings

We used multidimensional recurrence quantification analysis (MdRQA) to identify recurrent behavioral patterns during a CPS task in the context of an educational physics game. We recorded speech rate, body movement, and galvanic skin response (GSR) from individuals in a triad as measures of verbal contributions to the collaboration, nonverbal communication, and autonomic arousal, respectively. The findings demonstrate that teams exhibit collective patterns of regularity that differ significantly from baselines, and the difference was larger at short time intervals. Further, we found that overall team activity in each modality was inversely related to regularity. That is, less overall activity in a modality corresponded to fewer unique values and more repeat states, in turn, contributing to a higher recurrence rate. Thus, MdRQA successfully identified systematic patterns in team-level dynamics spanning verbal and nonverbal behaviors as well as autonomic arousal.

We found that patterns of regularity were robust to contextual changes during the CPS task, including assigned problem solving goal, physics concept, and experimental block (first or second collaborative interaction). It is possible that teams are able to maintain structure in their interactions despite perturbations to the task. It could also be the case that a more significant task perturbation is necessary for teams to reorganize their interaction processes, a hypothesis which needs further experimentation to validate.

We also found trends in regularity within each team interaction. Teams displayed an initial increase and subsequent decrease in regularity during the 15-minute interactions. It is possible that teams began with a highly irregular behavior as they became acquainted with the new task goals. As the interaction progressed, teams' increase in regularity may be attributed to the establishment of common ground [7]. During the second half of the interaction, teams may begin to reform their dynamics in an attempt to try new solutions and complete the task under the time constraints. In particular, the reminder indicating there were five minutes left in the task could perturb stable behavioral patterns that the team had previously settled into. It is likely that teams engaged in more irregular behavior as they attempted to complete the task before time ran out.

Finally, we found that regularity of our multimodal signals was predictive of CPS outcomes. Although regularity was not related to self-reported arousal or the perceived quality of the collaboration, it predicted valence. This suggests that teams with more repetitive patterns felt that the collaboration was more unpleasant. We also found that irregularity predicted task performance under certain task contexts. Specifically, the fact that regularity only predicted performance for the easier properties of torque concept but not the energy transfer concept provides a boundary condition for this effect.

Although it may be surprising that irregularity predicted higher performance, the most similar research [3] on a different dataset has linked irregularity to an increase in effective CPS processes, such as shared knowledge construction and negotiation/coordination, which supports our findings. Our basic interpretation of this finding is that teams exhibiting less regularity, which indexes more novel and less deterministic

behavioral configurations, were better able to adapt to the challenging task. This metric of behavioral irregularity might be an exciting new measure to index team performance in some collaborative contexts.

5.2 Applications

Our findings could be used to inform intelligent systems that monitor the quality of ongoing interactions and intervene when necessary. In particular, in the context of virtual interaction, regularity of low-level behaviors could be monitored in real-time in video conferencing software. Targeted interventions could be delivered during the collaboration as needed. For example, prolonged states of heightened regularity may signify stagnation. Thus, the system could encourage teams to try new solutions or communicate in a different way. In addition to real-time feedback, these findings could be applied to post-collaboration assessment and feedback about collaborative processes. The recurrence plots themselves can provide qualitative information about when teams engage in regular versus irregular behaviors and can become a core part of interactive visualization software that enables teams to zoom in on periods of interest and even search for specific recurrent behaviors (e.g., periods of low arousal and body movement but a lot of verbal communication). Since automated interaction feedback systems have been tested and proven beneficial to team performance and behavior [49, 77, 88], it is likely that feedback with our metrics could also effectively guide productive collaboration.

Measures of regularity could be used to aid machine learning models in predicting collaborative processes and outcomes. For instance, recurrence rate could be calculated over short time intervals and used as an input feature along with other multimodal features. This might provide a boost in machine learning performance over more primitive features because it indexes system-level dynamics comprised from multiple interaction modalities and interacting individuals.

5.3 Limitations and future work

Our work has five main limitations that should be addressed in the future. First, our subject sample consists of students at two universities and is not demographically representative of the general population, particularly with respect to age diversity. Second, this study was conducted in a controlled lab environment. Interaction dynamics will presumably change in-the-wild when there are increased noise and distractions. Third, while we did assess changing task contexts (i.e., goals and physics concept manipulation), the analyses were conducted on single task environment (i.e., a physics game), which limits generalizability claims. Fourth, we examined short, 15-minute interactions in teams where the individuals were mostly unfamiliar with each other. Interaction patterns might change as teams interact for longer time periods or familiarity increases. Fifth, this work can be expanded to include signals like facial expression and acoustic-prosodic features, to index emotional aspects of interaction, eye gaze to index attention, and language to index the content of the interaction. We are currently analyzing these additional signals across multiple tasks, with the goal of understanding interaction dynamics from a more holistic perspective (i.e., verbal and nonverbal communication, emotion, physiology, and attention). We expect these additional channels will increase our understanding of complex multimodal interactive patterns and explain more variance in predicting collaborative outcomes, a hypothesis that warrants future research.

5.4 Concluding remarks

We investigated multimodal regularity in speech rate, body movement, and galvanic skin response as triads engaged in a collaborative problem solving task. We linked regularity patterns to subjective and objective task outcomes in some contexts. Thus, multimodal regularity might reflect a promising approach to study team-level dynamics in multiparty interactions.

ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation (NSF DUE 1745442) and the Institute of Educational Sciences (IES R305A170432). 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 funding agencies.

REFERENCES

[1] Alterman, R. and Harsch, K. 2017. A more reflective form of joint problem solving. *International Journal of Computer-Supported Collaborative Learning*. 12, 1 (Mar. 2017), 9–33. DOI:https://doi.org/10.1007/s11412-017-9250-1.

[2] Ambady, N. and Rosenthal, R. 1993. Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*. 64, 3 (1993), 431.

[3] Amon, M.J., Vrzakova, H. and D'Mello, S.K. 2019. Beyond dyadic synchrony: Multimodal behavioral irregularity predicts quality of triadic collaborative problem solving. *Cognitive Science*. (2019).

[4] Aran, O. and Gatica-Perez, D. 2013. One of a kind. *Proceedings of the 15th ACM on International conference on multimodal interaction - ICMI '13* (New York, New York, USA, 2013), 11–18.

[5] Ashenfelter, K.T. 2008. Simultaneous analysis of verbal and nonverbal data during conversation: Symmetry and turn-taking. *Dissertation Abstracts International, B: Sciences and Engineering*. (2008).

[6] Avci, U. and Aran, O. 2016. Predicting the performance in decision-making tasks: From individual cues to group interaction. *IEEE Transactions on Multimedia*. 18, 4 (Apr. 2016), 643–658. DOI:https://doi.org/10.1109/TMM.2016.2521348.

[7] Barron, B. 2000. Achieving coordination in collaborative problem-solving groups. *Journal of the Learning Sciences*. 9, 4 (Oct. 2000), 403–436. DOI:https://doi.org/10.1207/S15327809JLS0904_2.

[8] Barsade, S.G. and Gibson, D.E. 2012. Group affect. *Current Directions in Psychological Science*. 21, 2 (Apr. 2012), 119–123. DOI:https://doi.org/10.1177/0963721412438352.

[9] Basu, S., Choudhury, T., Clarkson, B. and Pentland, A. 2001. Towards measuring human interactions in conversational settings. *Proc. IEEE CVPR Workshop on Cues in Communication* (2001).

[10] Bates, D., Mächler, M., Bolker, B. and Walker, S. 2015. Fitting linear mixed-effects models using {lme4}. *Journal of Statistical Software*. 67, 1 (2015), 1–48. DOI:https://doi.org/10.18637/jss.v067.i01.

[11] Beyan, C., Carissimi, N., Capozzi, F., Vascon, S., Bustreo, M., Pierro, A., Becchio, C. and Murino, V. 2016. Detecting emergent leader in a meeting environment using nonverbal visual features only. *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016* (New York, New York, USA, 2016), 317–324.

[12] Binkley, M., Erstad O., H., J. Raizen, S., Ripley, M., Miller-Ricci, M. and Rumble, M. 2012. Defining twenty-first century skills. *Assessment and teaching of 21st century skills*.

[13] Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., Wang, L. and Zhao, W. 2015. Automatic detection of learning-centered affective states in the wild. *Proceedings of the 20th international conference on intelligent user interfaces* (2015), 379–388.

[14] Braithwaite, J., Watson, D., Robert, J. and Mickey, R. 2013. A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiological Research*. (2013).

[15] Brannick, M.T. and Prince, C. 1997. An overview of team performance management. *Team performance assessment and measurement: theory, methods, and applications*.

[16] Burleson, B.R. and Greene, J.O. 2003. *Handbook of communication and social interaction skills*. L. Erlbaum Associates.

[17] Butler, E.A. and Randall, A.K. 2013. Emotional Coregulation in Close Relationships. *Emotion Review*. 5, 2 (Apr. 2013), 202–210. DOI:https://doi.org/10.1177/1754073912451630.

[18] Coco, M.I. and Dale, R. 2014. Cross-recurrence quantification analysis of categorical and continuous time series: an R package. *Frontiers in Psychology*. 5, (Jun. 2014). DOI:https://doi.org/10.3389/fpsyg.2014.00510.

[19] Conley, D.T. 2014. The common core state standards: Insight into their development. Washington, DC: Council of Chief State School Officers.

[20] Cornejo, C., Cuadros, Z., Morales, R. and Paredes, J. 2017. Interpersonal coordination: Methods, achievements, and challenges. *Frontiers in Psychology*. 8, (Sep. 2017).

DOI:https://doi.org/10.3389/fpsyg.2017.01685.
[21] Critchley, H.D., Elliott, R., Mathias, C.J. and Dolan, R.J. 2000. Neural activity relating to generation and representation of galvanic skin conductance responses: a functional magnetic resonance imaging study. *Journal of Neuroscience*. 20, 8 (2000), 3033–3040.
[22] D'Mello, S., Dale, R. and Graesser, A. 2012. Disequilibrium in the mind, disharmony in the body. *Cognition & Emotion*. 26, 2 (2012), 362–374.
[23] Dale, R., Fusaroli, R., Duran, N.D. and Richardson, D.C. 2013. The self-organization of human interaction. *Psychology of Learning and Motivation*. Elsevier. 43–95.
[24] Dale, R., Warlaumont, A.S. and Richardson, D.C. 2011. Nominal cross recurrence as a generalized lag sequential analysis for behavioral streams. *International Journal of Bifurcation and Chaos*. 21, 04 (2011), 1153–1161.
[25] Ekman, P., Levenson, R. and Friesen, W. 1983. Autonomic nervous system activity distinguishes among emotions. *Science*. 221, 4616 (Sep. 1983), 1208–1210. DOI:https://doi.org/10.1126/science.6612338.
[26] Ekman, R. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.
[27] Evans, C.R. and Dion, K.L. 1991. Group cohesion and performance. *Small Group Research*. 22, 2 (May 1991), 175–186. DOI:https://doi.org/10.1177/1046496491222002.
[28] Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., Andre, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S. and Truong, K.P. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*. 7, 2 (Apr. 2016), 190–202. DOI:https://doi.org/10.1109/TAFFC.2015.2457417.
[29] Flor, M., Yoon, S.-Y., Hao, J., Liu, L. and von Davier, A. 2016. Automated classification of collaborative problem solving interactions in simulated science tasks. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (2016), 31–41.
[30] Gardner, D.G. and Pierce, J.L. 2016. Organization-based self-esteem in work teams. *Group Processes & Intergroup Relations*. 19, 3 (May 2016), 394–408. DOI:https://doi.org/10.1177/1368430215590491.
[31] Gigone, D. and Hastie, R. 1993. The common knowledge effect: Information sharing and group judgment. *Journal of Personality and Social Psychology*. 65, 5 (1993), 959–974. DOI:https://doi.org/10.1037/0022-3514.65.5.959.
[32] Gil, M.C. and Henning, R.A. 2000. Determinants of perceived teamwork: Examination of team performance and social psychophysiology. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 44, 12 (Jul. 2000), 2-743-2–746. DOI:https://doi.org/10.1177/154193120004401283.
[33] Gorman, J.C., Amazeen, P.G. and Cooke, N.J. 2010. Team coordination dynamics. *Research in Higher Education*. (2010).
[34] Gorman, J.C., Cooke, N.J., Amazeen, P.G. and Fouse, S. 2012. Measuring patterns in team interaction sequences using a discrete recurrence approach. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. 54, 4 (Aug. 2012), 503–517. DOI:https://doi.org/10.1177/0018720811426140.
[35] Grafsgaard, J., Duran, N., Randall, A., Tao, C. and D'Mello, S. 2018. Generative multimodal models of nonverbal synchrony in close relationships. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (2018), 195–202.
[36] Griffin, P., Care, E. and McGaw, B. 2012. The changing role of education and schools. *Assessment and teaching of 21st century skills*. Springer. 1–15.
[37] Gully, S.M., Devine, D.J. and Whitney, D.J. 1995. A meta-analysis of cohesion and performance. *Small Group Research*. 26, 4 (Nov. 1995), 497–520. DOI:https://doi.org/10.1177/1046496495264003.
[38] Häkkinen, P., Järvelä, S., Mäkitalo-Siegl, K., Ahonen, A., Näykki, P. and Valtonen, T. 2017. Preparing teacher-students for twenty-first-century learning practices (PREP 21): a framework for enhancing collaborative problem-solving and strategic learning skills. *Teachers and Teaching*. 23, 1 (Jan. 2017), 25–41. DOI:https://doi.org/10.1080/13540602.2016.1203772.
[39] Harrison, D.A., Price, K.H. and Bell, M.P. 1998. Beyond relational demography: Time and the effects of surface- and deep-level diversity on work group cohesion. *Academy of Management*

Journal. 41, 1 (Feb. 1998), 96–107. DOI:https://doi.org/10.2307/256901.

[40] Helm, J.L., Sbarra, D. and Ferrer, E. 2012. Assessing cross-partner associations in physiological responses via coupled oscillator models. *Emotion*. 12, 4 (2012), 748–762. DOI:https://doi.org/10.1037/a0025036.

[41] Henning, R.A., Armstead, A.G. and Ferris, J.K. 2009. Social psychophysiological compliance in a four-person research team. *Applied Ergonomics*. 40, 6 (Nov. 2009), 1004–1010. DOI:https://doi.org/10.1016/j.apergo.2009.04.009.

[42] Henning, R.A. and Korbelak, K.T. 2005. Social-psychophysiological compliance as a predictor of future team performance. *Psychologia*. 48, 2 (2005), 84–92.

[43] Hesse, F., Care, E., Buder, J., Sassenberg, K. and Griffin, P. 2015. A framework for teachable collaborative problem solving skills. *Assessment and teaching of 21st century skills*. Springer. 37–56.

[44] Ishii, R., Otsuka, K., Kumano, S., Higashinaka, R. and Tomita, J. 2018. Analyzing gaze behavior and dialogue act during turn-taking for estimating empathy skill level. *Proceedings of the 2018 on International Conference on Multimodal Interaction - ICMI '18* (New York, New York, USA, 2018), 31–39.

[45] Jayagopi, D.B., Hung, H., Yeo, C. and Gatica-Perez, D. 2009. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing*. 17, 3 (Mar. 2009), 501–513. DOI:https://doi.org/10.1109/TASL.2008.2008238.

[46] Jermann, P., Mullins, D., Nüssli, M.-A. and Dillenbourg, P. 2011. Collaborative gaze footprints: Correlates of interaction quality. *International Conference on Computer Supported Collaborative Learning* (2011).

[47] Kendon, A. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

[48] Kerr, N.L. and Tindale, R.S. 2004. Group performance and decision making. *Annual Review of Psychology*. 55, 1 (Feb. 2004), 623–655. DOI:https://doi.org/10.1146/annurev.psych.55.090902.142009.

[49] Kim, T., Chang, A., Holland, L. and Pentland, A.S. 2008. Meeting mediator: enhancing group collaborationusing sociometric feedback. *Proceedings of the 2008 ACM conference on Computer supported cooperative work* (2008), 457–466.

[50] Konvalinka, I., Xygalatas, D., Bulbulia, J., Schjodt, U., Jegindo, E.-M., Wallot, S., Van Orden, G. and Roepstorff, A. 2011. Synchronized arousal between performers and related spectators in a fire-walking ritual. *Proceedings of the National Academy of Sciences*. 108, 20 (May 2011), 8514–8519. DOI:https://doi.org/10.1073/pnas.1016955108.

[51] Ku, H.-Y., Tseng, H.W. and Akarasriworn, C. 2013. Collaboration factors, teamwork satisfaction, and student attitudes toward online collaborative learning. *Computers in Human Behavior*. 29, 3 (May 2013), 922–929. DOI:https://doi.org/10.1016/j.chb.2012.12.019.

[52] Langfred, C.W. 1998. Is group cohesiveness a double-edged sword? *Small Group Research*. 29, 1 (Feb. 1998), 124–143. DOI:https://doi.org/10.1177/1046496498291005.

[53] Latif, N., Barbosa, A. V., Vatiokiotis-Bateson, E., Castelhano, M.S. and Munhall, K.G. 2014. Movement coordination during conversation. *PLoS ONE*. 9, 8 (Aug. 2014), e105036. DOI:https://doi.org/10.1371/journal.pone.0105036.

[54] Laughlin, P.R. and Ellis, A.L. 1986. Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*. 22, 3 (1986), 177–189.

[55] Laughlin, P.R., Hatch, E.C., Silver, J.S. and Boh, L. 2006. Groups perform better than the best individuals on letters-to-numbers problems: Effects of group size. *Journal of Personality and social Psychology*. 90, 4 (2006), 644.

[56] Laughlin, P.R., Kerr, N.L., Davis, J.H., Halff, H.M. and Marciniak, K.A. 1975. Group size, member ability, and social decision schemes on an intellective task. *Journal of Personality and Social Psychology*. 31, 3 (1975), 522.

[57] Levine, J.M. 2008. *Small groups*. Psychology Press.

[58] Marwan, N., Romano, C.M., Thiel, M. and Kurths, J. 2007. Recurrence plots for the analysis of complex systems. *Physics Reports*. 438, 5–6 (Jan. 2007), 237–329. DOI:https://doi.org/10.1016/j.physrep.2006.11.001.

[59] McManus, M.M. and Aiken, R.M. 2016. Supporting effective collaboration: Using a rearview mirror to look forward. *International Journal of Artificial Intelligence in Education*. 26, 1 (Mar. 2016), 365–377. DOI:https://doi.org/10.1007/s40593-015-0068-6.

[60] Miguel, J., Andres, L., Mercedes, M., Rodrigo, T. and Sugay, J.O. 2014. An exploratory analysis of confusion among students using Newton's Playground. *Proceedings of the 22nd International Conference on Computers in Education*. (2014).

[61] Morrow, L., Vrtunski, P.B., Kim, Y. and Boller, F. 1981. Arousal responses to emotional stimuli and laterality of lesion. *Neuropsychologia*. 19, 1 (1981), 65–71.

[62] Murray, G. and Oertel, C. 2018. Predicting group performance in task-based interaction. *Proceedings of the 2018 on International Conference on Multimodal Interaction* (2018), 14–20.

[63] Nijstad, B.A., Stroebe, W. and Lodewijkx, H.F.M. 2003. Production blocking and idea generation: Does blocking interfere with cognitive processes? *Journal of Experimental Social Psychology*. 39, 6 (Nov. 2003), 531–548. DOI:https://doi.org/10.1016/S0022-1031(03)00040-4.

[64] Nüssli, M.-A. and Jermann, P. 2012. Effects of sharing text selections on gaze cross-recurrence and interaction quality in a pair programming task. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12* (New York, New York, USA, 2012), 1125.

[65] Okada, S., Ohtake, Y., Nakano, Y.I., Hayashi, Y., Huang, H.-H., Takase, Y. and Nitta, K. 2016. Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets. *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016* (New York, New York, USA, 2016), 169–176.

[66] Palumbo, R. V., Marraccini, M.E., Weyandt, L.L., Wilder-Smith, O., McGee, H.A., Liu, S. and Goodwin, M.S. 2017. Interpersonal autonomic physiology: A systematic review of the literature. *Personality and Social Psychology Review*. 21, 2 (May 2017), 99–141. DOI:https://doi.org/10.1177/1088868316628405.

[67] Peñarroja, V., Orengo, V. and Zornoza, A. 2017. Reducing perceived social loafing in virtual teams: The effect of team feedback with guided reflexivity. *Journal of Applied Social Psychology*. 47, 8 (Aug. 2017), 424–435. DOI:https://doi.org/10.1111/jasp.12449.

[68] PISA, O. 2010. Field trial problem solving framework: Draft subject to possible revision after the field trial. Paris: OECD.

[69] PISA, O. 2015. Field trial problem solving framework: Draft subject to possible revision after the field trial. Paris: OECD.

[70] Rendeiro, C., Vauzour, D., Kean, R.J., Butler, L.T., Rattray, M., Spencer, J.P.E. and Williams, C.M. 2012. Blueberry supplementation induces spatial memory improvements and region-specific regulation of hippocampal BDNF mRNA expression in young rats. *Psychopharmacology*. 223, 3 (Oct. 2012), 319–330. DOI:https://doi.org/10.1007/s00213-012-2719-8.

[71] Richardson, D.C. and Dale, R. 2005. Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*. 29, 6 (Nov. 2005), 1045–1060. DOI:https://doi.org/10.1207/s15516709cog0000_29.

[72] Richardson, D.C., Dale, R. and Kirkham, N.Z. 2007. The art of conversation is coordination. *Psychological science*. 18, 5 (2007), 407–413.

[73] Riley, M.A., Richardson, M.J., Shockley, K. and Ramenzoni, V.C. 2011. Interpersonal synergies. *Frontiers in Psychology*. 2, (2011). DOI:https://doi.org/10.3389/fpsyg.2011.00038.

[74] Roschelle, J. and Teasley, S.D. 1995. The construction of shared knowledge in collaborative problem solving. *Computer supported collaborative learning* (1995), 69–97.

[75] Rosen, Y. 2015. Computer-based assessment of collaborative problem solving: Exploring the feasibility of human-to-agent approach. *International Journal of Artificial Intelligence in Education*. 25, 3 (2015), 380–406.

[76] Rosen, Y. and Rimor, R. 2016. Teaching and assessing problem solving in online collaborative environment. *Professional Development and Workplace Learning: Concepts, Methodologies, Tools, and Applications*. IGI Global. 254–269.

[77] Samrose, S., Zhao, R., White, J., Li, V., Nova, L., Lu, Y., Ali, M.R. and Hoque, M.E. 2018. CoCo: Collaboration Coach for understanding team dynamics during video conferencing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. 1, 4 (2018), 160.

[78] Sanchez-Cortes, D., Aran, O., Jayagopi, D.B., Schmid Mast, M. and Gatica-Perez, D. 2013. Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces*. 7, 1–2 (Mar. 2013), 39–53. DOI:https://doi.org/10.1007/s12193-012-0101-0.

[79] Schmidt, R.C. and Richardson, M.J. 2008. Dynamics of interpersonal coordination. *Coordination: Neural, behavioral and social dynamics*. Springer. 281–308.

[80] Schulze, J. and Krumm, S. 2017. The “virtual team player”: A review and initial model of knowledge, skills, abilities, and other characteristics for virtual collaboration. *Organizational Psychology Review*. (2017). DOI:https://doi.org/10.1177/2041386616675522.

[81] Sears, D.A. and Reagin, J.M. 2013. Individual versus collaborative problem solving: divergent outcomes depending on task complexity. *Instructional Science*. 41, 6 (Nov. 2013), 1153–1172. DOI:https://doi.org/10.1007/s11251-013-9271-8.

[82] Shimmer Sensing 2017. GSR+ user guide.

[83] Shockley, K., Santana, M.-V. and Fowler, C.A. 2003. Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance*. 29, 2 (Apr. 2003), 326–332. DOI:https://doi.org/10.1037/0096-1523.29.2.326.

[84] Shockley, K.D. 2005. Cross recurrence quantification of interpersonal postural activity. *Tutorials in contemporary nonlinear methods for the behavioral sciences*.

[85] signal developers 2014. {signal}: Signal processing.

[86] Spottiswoode, S.J.P. and May, E.C. 2003. Skin conductance prestimulus response: Analyses, artifacts and a pilot study. *J Scient Explor*. (2003).

[87] Stewart, A.E.B., Keirn, Z.A. and D’Mello, S.K. 2018. Multimodal modeling of coordination and coregulation patterns in speech rate during triadic collaborative problem solving. *Proceedings of the 2018 on International Conference on Multimodal Interaction* (2018), 21–30.

[88] Tausczik, Y.R. and Pennebaker, J.W. 2013. Improving teamwork using real-time language feedback. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2013), 459–468.

[89] Vaidyanathan, P., Pelz, J., Alm, C., Shi, P. and Haake, A. 2014. Recurrence quantification analysis reveals eye-movement behavior differences between experts and novices. *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA ’14* (New York, New York, USA, 2014), 303–306.

[90] Voyles, E.C., Bailey, S.F. and Durik, A.M. 2015. New pieces of the jigsaw classroom: Increasing accountability to reduce social loafing in student group projects. *The New School Psychology Bulletin*. (2015).

[91] Vrzakova, H., Stewart, A.E.B. and Amon, M.J. 2019. Dynamics of visual attention in multiparty collaborative problem solving using multidimensional recurrence quantification analysis. (2019), 1–14.

[92] Wallot, S., Roepstorff, A. and Mønster, D. 2016. Multidimensional Recurrence Quantification Analysis (MdRQA) for the analysis of multidimensional time-series: A software implementation in MATLAB and its application to group-level data in joint action. *Frontiers in Psychology*. 7, (Nov. 2016). DOI:https://doi.org/10.3389/fpsyg.2016.01835.

[93] Webber Jr, C.L. and Zbilut, J.P. 2005. Recurrence quantification analysis of nonlinear dynamical systems. *Tutorials in contemporary nonlinear methods for the behavioral sciences*. 94, (2005), 26–94.

[94] von Zimmermann, J. and Richardson, D.C. 2016. Verbal synchrony and action dynamics in large groups. *Frontiers in Psychology*. 7, (Dec. 2016). DOI:https://doi.org/10.3389/fpsyg.2016.02034.