# Recurrence Quantification Analysis of Eye Gaze Dynamics During Team Collaboration

ROBERT G. MOULDER, University of Colorado Boulder, USA
BRANDON M. BOOTH, University of Colorado Boulder, USA
ANGELINA ABITINO, University of Colorado Boulder, USA
SIDNEY K. D'MELLO, University of Colorado Boulder, USA

Shared visual attention between team members facilitates collaborative problem solving (CPS), but little is known about how team-level eye gaze dynamics influence the quality and successfulness of CPS. To better understand the role of shared visual attention during CPS, we collected eye gaze data from 279 individuals solving computer-based physics puzzles while in teams of three. We converted eye gaze into discrete screen locations and quantified team-level gaze dynamics using recurrence quantification analysis (RQA). Specifically, we used a centroid-based auto-RQA approach, a pairwise team member cross-RQAs approach, and a multi-dimensional RQA approach to quantify team-level eye gaze dynamics from the eye gaze data of team members. We find that teams differing in composition based on prior task knowledge, gender, and race show few differences in team-level eye gaze dynamics. We also find that RQA metrics of team-level eye gaze dynamics were predictive of task success (all *p*s<.001). However, the same metrics showed different patterns of feature importance depending on predictive model and RQA type, suggesting some redundancy in task-relevant information. These findings signify that team-level eye gaze dynamics play an important role in CPS and that different forms of RQA pick up on unique aspects of shared attention between team-members.

CCS Concepts: • **Human-centered computing** → **Collaborative interaction**; **Empirical studies in collaborative and social computing**; • **Applied computing** → *Psychology*.

Additional Key Words and Phrases: recurrence quantification analysis, team dynamics, eye gaze dynamics, team collaboration, shared attention

**ACM Reference Format:**
Robert G. Moulder, Brandon M. Booth, Angelina Abitino, and Sidney K. D'Mello. 2023. Recurrence Quantification Analysis of Eye Gaze Dynamics During Team Collaboration. In *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023), March 13–17, 2023, Arlington, TX, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3576050.3576113

## 1 INTRODUCTION

Humanity's problems, big or small, are rarely solved by singular individuals. Instead, groups of individuals with diverse backgrounds and complementary skill sets work together to solve complex problems. For instance, elementary school students may form groups to complete a large art project, college students may work on teams when conducting a hazardous chemistry experiment, or a team of software engineers may work together to build a new application. Unfortunately, only a minority of individuals exhibit high levels of collaboration proficiency [40], and this scarcity has been further exacerbated by the recent increase in remote collaboration as a result of the COVID-19 pandemic [4, 20]. In particular, individuals report that new cognitive challenges (e.g., time management, learning new technologies, and adapting to rapid/inconsistent changes) and social challenges (e.g., feelings of loneliness, a lack of motivation


Manuscript submitted to ACM

or engagement, and difficulty acquiring personalized assistance) significantly impact their collaborative experiences [19, 33].

Research in collaborative problem solving (CPS), where two or more individuals work together to construct and execute in problem solving, aims to understand the conditions under which team efforts result in successful outcomes [2]. Successful CPS requires that team members exhibit both the cognitive skills (e.g., planning, reasoning) as well as the social skills (e.g., negotiation, conflict resolution) needed to create effective solutions to complex problems [40]. A related and emerging research field called collaborative analytics aims to characterize relevant aspects of CPS (e.g., team makeup, collaboration success, and collaborative learning) using multimodal data streams from multiparty interactions[7, 15, 24], with the end-goal of incorporating CPS measures into feedback technologies that aim to improve collaboration processes and outcomes while attenuating barriers to successful CPS. Work in this area has been successful in using numerous combinations of data modalities such as speech [17, 31, 34], facial expressions [44, 48], body movement [5, 32, 48], physiological indicators [13], and eye gaze [30, 48] to predict certain CPS processes and outcomes, and it has produced supportive CPS technologies. Examples include real-time intelligent team tutoring systems (ITTs) where a computer interacts with multiple team members synchronously [41], AI teammates that can assist in regulating group dynamics [12], or an AI system that can monitor CPS as it unfolds and provide feedback to group members about their collaboration skills after task completion [6, 42]. Many of these technologies analyze CPS through the lens of individual contributions and actions, but in highly collaborative teamwork, successful outcomes are often greater than the sum of individual contributions. Thus, in this work we adopt the view that group collaboration needs to be studied holistically at the level of teams rather than individuals.

One major open research question is how to analyze signals from multiple individuals for statistical or computational modeling of team collaboration and how teams progress towards successful outcomes [41]. Although using a combination of modalities has shown improvement from unimodal or bimodal models in modeling CPS [47], few works have focused on the multiparty aspect examining how complex patterns of individual and team actions lead to success. Therefore, this work focuses on improving the multiparty element instead of the multimodal aspect of modeling collaborative interactions.

We situate our multiparty analysis in the context of triads engaged in active remote problem solving with a shared screen, and we target a single modality: eye gaze. We choose eye gaze since it indexes social visual attention [11], captures elements of cognitive state [36], and since the dynamics and coordination of individuals' gaze within a team setting are inherently complex and nonlinear [35, 51]. As a novel solution to account for the inherent complexity of eye gaze dynamics in multiparty interactions during CPS, we propose a using an approach derived from nonlinear dynamical systems theory known as recurrence quantification analysis (RQA) for processing multiparty gaze signals during a group CPS task. RQA is a method of quantifying both linear and nonlinear dynamics of time series by constructing a symbolic representation of the underlying dynamics observed from these time series. RQA has multiple variants, each applicable to different situations, which we examine in this work. Specifically, we study how auto-recurrence quantification analysis (ARQA; a method for quantifying the internal dynamics of a singular time series), cross-recurrence quantification analysis (CRQA; a method for quantifying the shared dynamics of a two time series), and multi-dimensional recurrence quantification analysis (MdRQA; a method for quantifying the dynamics of repeating patterns across multiple time series) may be implemented for studying team-level eye gaze dynamics. In the remainder of this work, we describe each of these RQA methods in detail, discuss quantitative metrics that can be derived from RQA analysis, and we demonstrate their utility in capturing team eye gaze dynamics which are pertinent to inferring successful collaborative outcomes.

### 1.1 Background and Related Work

*1.1.1 Multimodal, Multiparty Modeling of CPS.* Using data from collaborative interactions to predict CPS processes and outcomes is a complex challenge that researchers have approached in a variety of ways. Much of the work in this area has explored combining multiple data modalities to robustly capture CPS dynamics. There is an extensive body of work on dyadic and multiparty interaction modeling to predict CPS outcomes (e.g., learning gains, collaboration quality, listener comprehension, group performance, task success, etc.) using combinations of modalities such as language [37, 39], facial expressions [44], body movements [5, 39], eye movements [34, 36] , electrodermal activity [34], speech and gaze cues (e.g., turn-taking, interruptions) [5], acoustic-prosodics [26, 44], and more. In this area, the focus has mainly been on how to combine multiple data modalities, and less on how to combine the signals from multiple individuals in a group, with the aim of capturing their dynamic interactions.

To address this gap, we direct our attention to modeling the multiparty elements of group interactions. In an effort to quantify these more complex multiparty group dynamics, some work has utilized nonlinear dynamic systems methods, but with an eye for analyzing the interaction instead of predicting CPS processes and outcomes [14]. In particular, Gorman et al. [14] combines nonlinear dynamics systems analysis (e.g., attractor reconstruction and Hurst exponent estimation) along with individual team member dynamics to model how team dynamics change when subject to outside perturbations. In line with Gorman et al. [14], other researchers have explored several techniques of combining data signals from multiple individuals [e.g., 23, 45].

One commonly utilized technique involves the concatenation of data signals across individuals to form a complete feature set for each team. This approach supports multimodal data capture for a variety of signals per individual, but depending on how features are concatenated, context regarding the intertwined group dynamics may be lost (e.g., group coordination and co-regulation) [8, 43]. Additionally, using an aggregative-statistic approach (i.e., computing high-level statistics from each individual's data) to derive group-level features may dilute the complex interplay between teammates [5]. Another approach involving strategically weighting individuals' signals based on characteristics, like role and behavior, found no improvement from an equal weighting baseline when predicting task success [44]. This suggests that an individual's assigned role during a collaborative task is not as influential as their effective role, which is not necessarily static since individual's often adapt to task needs; thus, models should account for this.

We extend the motivation for moving away from an aggregative-static approach for modeling multiparty systems to predicting CPS outcomes using predictive models rather than solely analyzing group behavior. To this end, we leverage recurrence quantification analysis as a means of capturing multiple levels of group dynamics and to account for individual teammates taking on shifting roles during a collaborative interaction, unlike dyadic conversational experiments which statically assign participants as either speaker or listener [36].

### 1.2 Recurrence Quantification Analysis

Recurrence quantification analysis is a nonlinear dynamical systems technique used to analyze recurrence, the repetition of patterns in a sequence [49, 50]. RQA has a long history of use within nonlinear dynamical systems research and it is used to recover/understand symbolic dynamics within a system [25]. Broadly, RQA can determine when a system returns to a previously visited state by comparing the similarity between time-series at every possible time lag.

The primary tool for doing this is the recurrence matrix (also known as a recurrence plot), a matrix where the x and y axes each represent a time-series, allowing for pair-wise comparison between every element (see Figure 1). Through this comparison, recurrence is identified as a "close" distance between time-series elements, where "close" is a user-defined

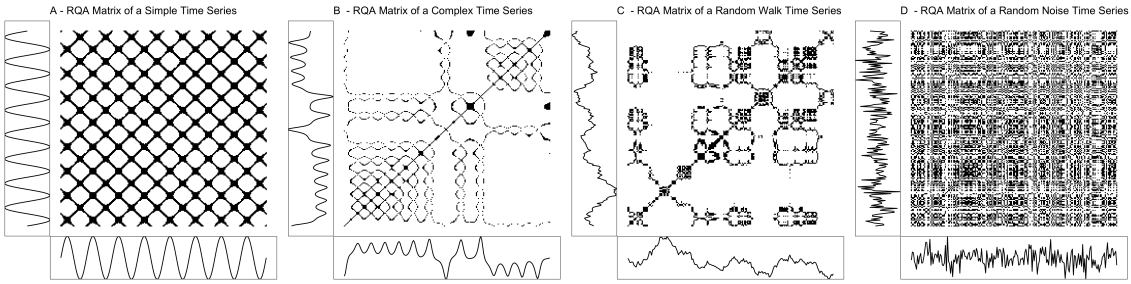


Fig. 1. Example of constructing recurrence quantification analysis (RQA) matrices from different time series. Patterns in the RQA matrices are indicative of intrinsic dynamics within each time series. (A) Represents RQA matrix construction of a simple periodic system; (B) Represents RQA matrix construction of a complex aperiodic system; (C) Represents RQA matrix construction of a structured random system (i.e., a random walk); (D) Represents RQA matrix construction of random noise. Each of these are example of auto-recurrence and are symmetric. Cross-recurrence plots are between two time series and are generally not symmetric.

parameter. RQA can also be generalized to using categorical time-series, making the distance comparison a binary decision (i.e., are the values being compared the same or not). This pairwise comparison of every time step produces the recurrence matrix, a binary matrix where 1 indicates recurrence and 0 indicates non-recurrence. The recurrence matrix itself is a powerful tool that can be used as a visual representation of the system's dynamics, displaying instances of regularity (repeated patterns), irregularity (lack of repeated patterns), and preserving temporal information.

In this work, we consider three approaches for performing RQA: auto-recurrence quantification analysis (ARQA), which specifically compares a time series to itself to determine whether there are repeated patterns within a signal, cross-recurrence quantification analysis (CRQA) which compares a single time series to a different time series to detect when the two signals are in sync or share common states, and multidimensional recurrence quantification analysis (MdRQA) which computes recurrence for more than two time series (see Figure 2).

**ARQA.** Auto-recurrence quantification analysis (ARQA) is a method for quantifying self-similar dynamics of a time series appearing at different time points across the time series. ARQA has been used in dyadic contexts to quantify physiological synchrony by performing ARQA on individuals' heart rate variability as an expression of socio-psychological compliance [46]. Since ARQA involves comparing a time series to itself, to be useful for studying team dynamics, it requires a singular team-level time series to be formed from some aggregation of individual team member dynamics.

In our context of eye gaze analysis, each individual's gaze locations with a team could be averaged to form a time series of centroid gaze locations. While this may seem an intuitive approach for ARQA analysis for studying team eye gaze dynamics, aggregation inherently destroys information that may be key to understanding eye gaze dynamics [8]. In the case of gaze location averaging, there is no information retained on the distance of each team member's gaze location to the centroid. Thus, times when all team members look at the same position on the screen could be treated the same as when all team members are looking at different extremes on the screen. This aggregation approach potentially damages the interpretability of team eye gaze dynamics derived from ARQA on aggregated time series, though the centroid gaze dynamics may still provide relevant information for understanding CPS in this context.

**Cross-RQA.** Cross-recurrence quantification analysis (CRQA) is a method of quantifying recurrent dynamics between pairs of different time series. CRQA is useful in dyadic settings to measure temporal coupling between individuals [9]. For example, CRQA has been used to identify that a lag time of 2-sec between a speaker and listener's eye movements

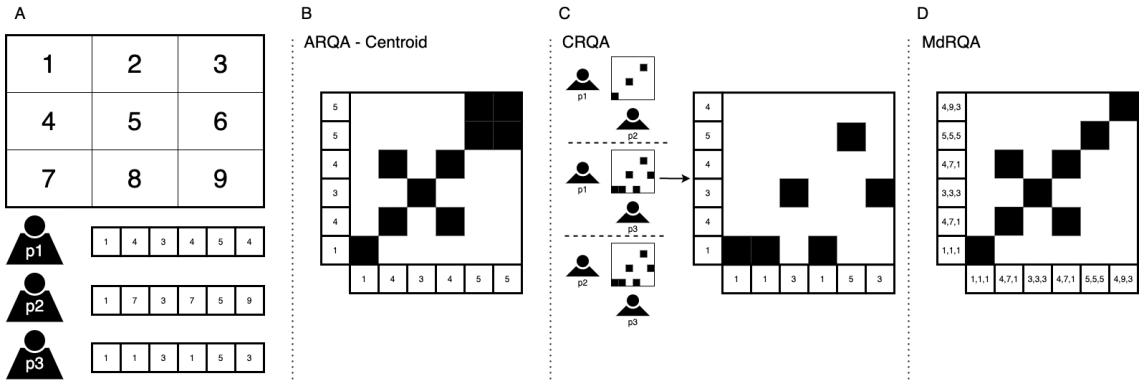


Fig. 2. A simplified demonstration of recurrence matrix formation from team gaze dynamics across a computer screen using different RQA approaches. (A) An example set of time series derived from participants gaze across a computer monitor. The monitor is broken up into nine regions. Three time series' of gaze locations from each participant (p1, p2, and p3) are recorded for six time points, which are then used to create RQA plots. (B) A time series of gaze centroids is calculated from all participant's gaze locations and an RQA plot is created relating the centroid time series to itself. (C) Gaze time series' from each pair of participants are used to create three unique recurrent matrices. (D) Multi-dimensional RQA (MdRQA) is conducted on all participant's time series simultaneously and points are considered recurrent when all three participants' gaze locations repeat at some other point in time.

maximizes the coupling between them [36]. Nüssli and Jermann [28] also used CRQA to capture joint attention as a measure of team synchrony by comparing gaze signals of dyads during a pair programming task while sharing text selections in order to manipulate levels of joint attention.

CRQA allows for a more natural and interpretable means of studying team-level gaze dynamics by utilizing metrics obtained from all individual pairs within a team. That is, by running a CRQA on every pair of team members, we may gain a better understanding of dyad-level dynamics, which can then be aggregated to the team-level. Compared to centroid-based ARQA, little information is lost during the aggregation process and team-level aggregated measures retain interpretability. However, because each CRQA matrix is based on the information of dyads only, it may be insufficient to quantify recurrent properties based for teams of three or more persons.

**Multi-dimensional RQA.** A third possible method for studying team dynamics in an RQA framework, and the only method requiring no aggregation, is MdRQA. Unlike centroid-based ARQA and CRQA, MdRQA matrices represent recurrence of the shared states of all participants simultaneously [49]. Amon et al. [2] previously used MdRQA to analyze speech rate, body movement, and user interface changes during a remote CPS task and found that irregularity (i.e., less recurrence indicating novel behaviors), an arguably neglected aspect of positive collaboration, predicted CPS processes related to shared knowledge and coordination. Additional work in this area, which also used MdRQA to process team signals, confirmed that team irregularity during more challenging tasks positively predicted task performance [13]. These works have utilized MdRQA to discover that irregularity is an influential factor during collaboration, demonstrating the power of analyzing recurrence across all participants simultaneously.

MdRQA provides a means of capturing the symbolic eye gaze dynamics of a team with minimal information loss. That is, if the gaze-location of all three participants at one time point is repeated at another time point, even if not in the same location, those points are considered recurrent in the MdRQA framework. This makes the interpretability of MdRQA the most direct of all of our studied RQA versions, in that all MdRQA metrics represent the repeated dynamics of gaze-fixation states between all team members at all time points. This also means that MdRQA and CRQA differ

slightly in their interpretations, as MdRQA represents the dynamics of repeated states between all team members simultaneously, while CRQA represents the dynamics of repeated states between pairs of team members.

### 1.3 Current Work: Contributions and Novelty

Our work aims to go beyond dyadic interactions by leveraging RQA techniques to analyze gaze signals of a triad. With RQA, we move ahead of just examining synchrony and joint attention, and instead move towards a team-level approach which examines how teams as a whole exhibit repeated patterns of behavior. Existing work on modeling multiparty signals has experimented with several strategies for modeling team-level dynamics such as concatenating feature sets for each individual, applying aggregative approaches which calculate statistics like the mean of all individual features, weighted averaging of individual features that express teammate differences (such as role and dominance), or pooling occurrences of behaviors across teammates [e.g. 23, 48]. However, there is a scarcity of work focusing on holistic multiparty analysis in the context of CPS.

In this study, we explore different multiparty modeling methods using recurrence analysis and provide findings that will benefit the learning analytics community in the context of group collaboration. To our knowledge, there is no other study that investigates the differences between different aggregation methods in an RQA framework for studying team gaze dynamics, only the gaze dynamics of individuals [3]. Determining an appropriate aggregation technique for understanding team-level dynamics from individual team member time series is an open question in regards to the dynamics of shared visual attention. Differing methods have their own interpretations, pros, and cons. As such, it is possible that the combination of information from each method may differentially contribute to accurately predicting CPS success. It is also possible that differences in team member composition may lead to differences in measured team dynamics derived from RQA. Therefore, we examine two research questions: **RQ1** - How does team composition (amount of prior knowledge, gender composition, and race composition) influence team dynamics as measured by RQA; and **RQ2** - How do recurrence matrix metrics integrating multiparty data and derived from different RQA methods differentially predict objective CPS task outcomes?

## 2 METHOD

### 2.1 Data

Data was obtained from an existing study on CPS [43]. All procedures were approved by each institute's respective Institutional Review Board and all participants provided written consent. Only aspects germane to the present study are discussed here (i.e., eye gaze dynamics, team composition, and team success outcomes).

Participants (N = 288; average age = 22; 56% female, 42% male, 2% other; 51% White, 27% Hispanic/Latino, 18% Asian, 3% Black, 1% reporting Native American or Other) were students from two large public universities in the Western US (111 from School 1 and 177 from School 2). Participants were assigned to 96 triads based on scheduling constraints. Each participant was compensated with a 50.00USD Amazon gift card (95.8%) or research credit (4.2%). Prior to the study, participants also completed a measure of knowledge of physics concepts. Data collection either occurred on computer-enabled workstations in separate rooms (School 1) or partitioned with dividers (School 2), to separate them from their team members in order to simulate a remote interaction. Workstations were fitted with webcams, headsets, and Tobii 4C eye trackers (for which licenses to record data were purchased). All interactions occurred over Zoom (https://zoom.us) with video and screen sharing. Gaze data was recorded at 90 Hz and is the primary data source analyzed here.

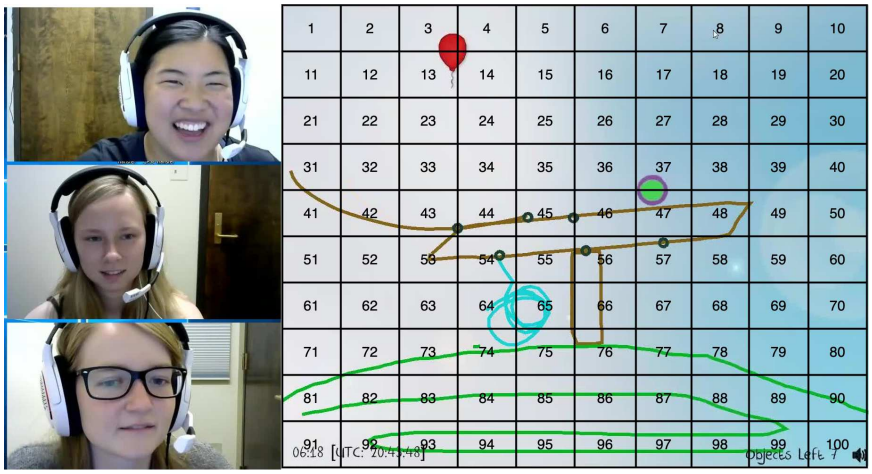


Fig. 3. Screenshot illustrating the 10x10 grid used to transform each individual's gaze location time-series into a categorical time series reflecting specific areas of the screen.

Triads were tasked with playing Physics Playground [1, 38], an educational computer game for learning Newtonian physics concepts. Teams participated in four 15-minute blocks, where the first three blocks involved the Physics Playground task, and the fourth block involved a task irrelevant to this study; therefore, we only used data from the first three blocks. The goal of Physics Playground is to guide a ball to a balloon, separated by obstacles, by drawing objects on the screen (e.g., ramps, levers, pendulums, springboards, weights); everything in the game obeys the laws of Newtonian physics. Teams earn a trophy when the goal is met, provided they do not exceed the maximum allowable number of objects per level. Physics Playground contains numerous levels featuring different obstacles, and teams were allowed to freely select which levels to attempt, when to quit a level to attempt another, and they were encouraged to complete as many levels as possible. For each block, one team member was assigned the role of *controller*. Their screen was shared with the rest of the team, and they were responsible for using their mouse to control the gameplay. The remaining teammates were labeled as *contributors* and able to communicate their ideas through audio and video, but they had no direct control over the game. The role of the controller was rotated between teammates each block so that each team member had the opportunity to be the controller one time. No level-specific support mechanisms were provided, except for a tutorial on general game mechanics and the user interface that could be viewed at any time.

### 2.2 Data Analytics

We collected the time series of gaze locations of all participants as (x, y) coordinates on the computer screen. We then created a gaze time-series for each participant by averaging gaze locations over fixed 225ms non-overlapping windows, as this was the mean fixation duration for our data set[1]. Next, we divided the computer screen (1920 x 1080 pixels) into a 10 x 10 grid to help measure gaze correspondence and numerically labeled each location. Then, we transformed the gaze time-series into categorical time-series by assigning each gaze location to a grid square from the screen (1-100) or off-screen (0) (see Figure 3).

[1]Fixations were obtained using PyGaze Analyser [10] and defined as frames in which gaze was maintained on a location (within 25 pixels) for at least 50ms and at most 1s.

The time-series were variable in length because they were generated separately from each level a team played, excluding the last 10s of each level to control for potential celebratory behaviors after earning a trophy. Each individual's categorical gaze time-series was associated with a binary measure of task success, where 0 was assigned if the individual's team did not earn a trophy for the level and 1 was assigned if they earned one (i.e., if they completed the level). In total, we generated time-series for 765 level attempts out of 789. Two teams' data were unavailable due to technical issues and one team's data had no valid gaze readings, leaving us with data from 93 teams. We imputed any invalid gaze data for these 93 teams using spline interpolation with a second-degree polynomial (approximately 7% missing frames on average).

*2.2.1 Recurrence Quantification Analysis of Multiparty Gaze Dynamics.* We calculated aggregated RQA metrics for understanding team-level gaze dynamics in three different ways (see Figure 2). First, we calculated a team-level centroid time series of gaze fixation locations on the 10x10 grid. This was done by taking the geometric center position of each participants gaze fixation location on the screen (using [x,y] coordinates) and then assigning a location on the 10x10 grid to each centroid value. When some team members gazed off the screen, only the gaze locations of the members focusing on the screen were used to create the time series of gaze centroids. When all team members were looking off screen, this was considered its own categorical state (state 0). This centroid time series was then assessed using ARQA. In an ARQA framework, a time series is compared against itself, resulting in a symmetric recurrence matrix (Figure 2.B). Second, we applied CRQA to each time series pair within each team and averaged the three resulting recurrence matrices from each pair. As opposed to ARQA, CRQA quantifies the recurrence between two separate time series, generally yielding a non-symmetric matrix (Figure 2.C). Finally, we applied MdRQA to all team members within a team simultaneously. MdRQA quantifies recurrence as the times when the categorized gaze locations of all team members at one point in time repeat at a later time (Figure 2.D).

For each RQA method, we utilized nine commonly used recurrence metrics to capture different characteristics of the corresponding recurrence matrix [25]. These features are: recurrence rate, determinism, diagonal line entropy, maximum diagonal line length, average diagonal line length, laminarity, vertical line entropy, maximum vertical line length, and average vertical line length. Measures involving diagonal lines capture evidence of regularity or repeated sequences of behavior, in our case, when one gaze location is likely to follow another. Measures involving vertical lines capture evidence of detailed inspection and reinspection of areas on the screen. A qualitative description of each feature and how each RQA feature relates specifically to eye gaze dynamics is provided below.

**Recurrence Rate.** *Recurrence rate* for categorical time series is simply the number of elements of a recurrence matrix that are 1 (black) out of all possible entries. That is, recurrence rate represents how often a system repeats itself. In the context of eye gaze dynamics during CPS, high recurrence rate is indicative of teams whose eye gaze patterns consistently visit/repeat previous states over time.

**Determinism, Average Diagonal Line Length, and Maximum Diagonal Line Length.** *Determinism* for categorical time series is calculated as the proportion of filled elements in a recurrence matrix that fall on diagonal lines. Whereas RQA matrices of random noise tend to have very few diagonal structures, more predictable and periodic time series tend to have regions with long diagonal lines (see Figure 1.A). Thus, diagonal lines in a recurrence matrix are indicative of the average predictability of shared change in dynamics of a system. Two common measures of determinism are the *average diagonal line length*, which measures the average length of time a system behaves in a manner similar to its previous states, and the *maximum diagonal line length* which quantifies the longest amount of time a system behaves in a manner similar to its previous states.

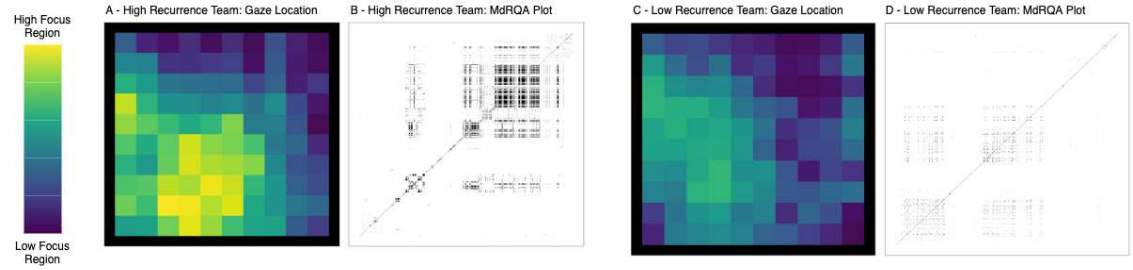


Fig. 4. Example team gaze dynamics across a screen and their associated MdRQA plots. (A) A heat map of gaze dynamics from a highly recurrent team. Darker blue colors indicate screen regions with low shared gaze across the study while brighter yellow colors indicate screen regions with high shared gaze. (B) The associated MdRQA plot for (A). (C) A heat map of gaze dynamics from a weakly recurrent team. (D) The associated MdRQA plot for (C).

**Laminarity, Average Vertical Line Length, and Maximum Vertical Line Length.** Similar to how determinism quantifies team dynamics using diagonal lines in an RQA matrix, *laminarity* for categorical time series is calculated as the proportion of filled elements in a recurrence matrix that fall on vertical lines. Vertical lines are indicative of intermittent phases of stability and instability within a system. In the context of team gaze analysis, when a snapshot of a team's gaze state at one moment repeats at several other points in time, the system will exhibit a high laminarity value [22] (see Figure 1.B). Similar to determinism, the *average vertical line length* and *maximum vertical line length* are indicative of the average length of time and maximum amount of time a time series shows laminar behavior respectively.

**Diagonal and Vertical Entropy.** A final set of RQA metrics we study in the paper revolve around the amounts of Shannon entropy[2] observed within the diagonal lines (determinism) and vertical lines (laminarity) within a recurrence matrix. *Diagonal entropy* is the calculated Shannon entropy for all possible diagonal line lengths observed in a recurrence matrix and thus captures uncertainty in the periodicity or predictability of a system. *Vertical entropy* is the Shannon entropy for all possible vertical line lengths observed in a recurrence matrix, and it captures the unpredictability of the system returning to some state.

To perform ARQA and CRQA we used the *crqa* R package [9]. To perform mdRQA for each team, we used the *mdrqa()* R function provided in [49] with multivariate time-series derived from grouping together the time-series from each member of a triad. Figure 4 shows example MdRQA recurrence matrices and gaze heat maps for teams exhibiting both high and low levels of gaze-dynamics recurrence.

### 2.3 Data Analysis: Team Composition and RQA Dynamics

The teams in this study were composed of individuals from differing backgrounds and skill sets. In order to determine if these demographic and experiential differences influenced our obtained RQA metrics of team dynamics (**RQ1**), we conducted a series of robust linear mixed-effects models,[3] predicting each RQA metric from each type of RQA from team average prior knowledge of physics concepts, team gender composition, and team race composition . These models take the form:

$$METRIC_{ij} = f(Prior_{0j}, Gender_{0j}, Race_{0j}) + u_{0j} + e_{ij} \quad (1)$$

[2]Shannon entropy is a measure of randomness or disorder within a systems. Shannon entropy is maximized when the distribution of a set of observations is uniformly distributed and is 0 when all observations share the same value.
[3]Linear mixed-effects models are a family of models for the analysis of nested data. The repeated-measures ANOVA/ANCOVA can be shown to a special case of a linear mixed-effects model, albeit with more strict assumptions. A robust model was chosen to account for the possibility of distributional assumption violations. All continuous variables were z-scored before inclusion into each model.

where $METRIC_{ij}$ is an RQA metric from either ARQA, CRQA, or MdRQA, for Physics Playground level $i$ within team $j$; $Prior_{0j}$ is the average prior knowledge of physics concepts for team $j$, $Gender_{0j}$ is a categorical variable representing if a team is all male (8%), all female (22%), or mixed gender (70%); $Race_{0j}$ is a categorical variable representing if a team is from all the same race (22%) or is racially diverse (78%); $u_{0j}$ is a random intercept term; and $e_{ij}$ is an error term.

### 2.4 Data Analysis: Trophy Prediction

To gauge the strength of the link between recurrence in teams' gaze patterns and successful team outcomes, we trained two machine learning models to predict successful level completion using different combinations of the RQA metrics and RQA approaches previously described (**RQ2**). The details of these machine learning experiments are described below.

**Data and Labels.** Each data sample corresponded to a team's level attempt and consisted of the team's RQA metrics (e.g., vertical entropy, average diagonal line length; 9 total) derived from each of the three RQA approaches (27 metrics in total) and a binary label corresponding to whether the level was successfully completed or not.

**Data Partitioning.** To facilitate using cross-validation to report model performance, we partitioned the data using team-independent stratified folds. First, the data samples were grouped by team (93 teams; approximately 8 level attempts per team) and then partitioned into three folds where each team's data was entirely contained within one fold. We employed a stratified sampling technique to ensure that the proportion of data samples in each fold corresponding to a successful level completion was about equal (approximately 53% level success rate per fold). This ensured that both successful and unsuccessful level attempts were present within each fold. This process was repeated 10 times using random initialization for the stratified sampling to produce 10 separate sets of three team-independent folds (i.e., 10 randomized cross-validation trials).

**Learning Algorithms.** Logistic regression and random forest learning models were selected for their interpretability. The logistic regression formulation included an $\ell_2$ regularization component to help prevent overfitting. All machine learning code was implemented using Python 3.6 and Scikit-learn 0.20.2 [29].

**Model Tuning.** Two hyperparameters for the random forest model were tuned using nested three-fold cross-validation with the same validation-set partitioning strategy as above. Per trial, during model training for each of the three folds, we further partitioned the training data to obtain validation folds and used a grid search to determine the optimal hyperparameters. The random forest was then retrained on all training data to predict the test data per fold. We tuned two parameters for the random forest: number of decision trees ({10, 15, 20, **25**, **50**, 100, 200, 300}), and the maximum tree depth ({**1**, **2**, 4, 6, 8, 10, 20, 30, 40}). We separately tuned random forest models for each set and combination of RQA metrics (see below), and the bold items indicate which values were chosen for at least one set of RQA features.

**Experiments.** For each of the logistic regression and random forest learning models and each of the RQA approaches (ARQA, CRQA, and MdRQA), we separately trained classifiers to predict whether a team completed a level. Additionally, we tested two model fusion approaches to investigate whether the RQA features contained mutual information relevant for predicting level success. In one approach (*feature fusion*), we combined all 9 features from each of the 3 RQA approaches (27 features in total) and then separately tuned and trained both the logistic regression and random forest models. In the second fusion approach (*late fusion*), we combined the predictions from the ARQA, CRQA, and MdRQA versions of the logistic regression model by computing their means (i.e., an ensemble), and likewise for the random forest *late fusion*. The predictions for teams' outcomes per level were evaluated using the area under the receiver operating characteristic curve (auROC).

**Shuffled Baseline.** To establish a fair basis for evaluating the utility of the RQA metrics, we generated *shuffled* variants for each of the above experiments. Each shuffled variant per experiment followed the same procedure, except just prior to model training, the level completion labels were randomly shuffled across all data samples. This preserved the base level success rate (53%), but mixed up these outcomes with respect to the RQA metrics, thus providing a simulated noise baseline for comparison.

## 3 RESULTS

### 3.1 Team Composition

In total, we ran 27 robust linear mixed-effects models for each combination of RQA approaches (ARQA, CRQA, MdRQA) and RQA metrics (9 total). For follow-up analyses, we used the R package *emmeans* and employed a false-discovery rate correction [21]. Out of all conducted tests, the only statistically significant difference in RQA metric by group composition variable was observed between all female teams and mixed-gender teams for *recurrence rate* of centroid-based ARQA ($p = .048$), indicating that all female teams ($Mean_{REC} = .07$) had slightly lower ARQA recurrence rate than mixed-gender teams ($Mean_{REC} = .08$). Thus, we consider the RQA analysis of CPS to be independent of individuals' prior knowledge and race while, it may exhibit a small differential effect across gender if ARQA metrics are utilized.

### 3.2 Predictive Accuracy and Feature Importance

Table 1 shows the auROC accuracy of logistic regression and random forest models for predicting whether teams successfully completed a level. The table shows the means and 95% confidence intervals for each model and RQA approach across the three cross-validation test folds for each of the 10 randomized trials (30 auROC samples in total). The ROC for each sample was generated from the test-set predictions within each cross-validation fold consisting of about 255 samples. Confidence intervals were estimated via stratified bootstrapping for each of the 30 ROCs individually and then averaged. Both the auROC and confidence intervals were computed using the *pROC* package in R.

Linear mixed-effects models were used to estimate the difference between the *original* and *shuffled* experiments. Each linear mixed-effects model was constructed to include a fixed effect comparing auROC values between *original* and *shuffled* RQA feature variants and a nested random effect accounting for paired samples within each cross-validation fold. Conceptually, this is similar to conducting a correlated paired-sample $t$-test, but it additionally accounts for paired-sample correlations in different cross-validation folds.

All *original* experiments significantly outperformed the *shuffled* model variants [e.g., $t(29.00) = 22.59, p < .001$, random forest late fusion model] and by extension also significantly outperformed random chance (e.g., auROC=.50) since none of the *shuffled* auROCs exceeded .50. This implies that our proposed RQA metrics of team gaze dynamics are indeed informative for predicting collaborative success in this context. Among the *original* logistic regression and random forest performances, random forest achieved higher auROCs than its corresponding logistic regression model, and the differences were significant for each set of RQA approaches ($p < .001$). This suggests that the link between RQA metrics and successful level completion is non-linear, which is unsurprising.

Within the random forest models for different RQA approaches, the differences in performance are small (<.05 auROC). Table 2 shows the correlations between the model predictions trained on RQA metrics from each of the three RQA approaches. All correlations are strong (>.60), however the predictions obtained from the CRQA and MdRQA metrics show the highest overall correlations for logistic regression and random forest (.84 in both cases), which suggests that CRQA and MdRQA features contain similar information. Thus, the results indicate, both in terms of auROCs and

Table 1. Comparison Between Mean (95% -CIs) Predictive Accuracy of Level Wins Between Original and Shuffled Variants

| | | auROC (N ≈ 255) | | |
|---|---|---|---|---|
| **Model** | **RQA Approach** | ***Original* Experiment** | ***Shuffled* Experiment** | ***Original* Vs. *Shuffled*** |
| Logistic Regression | ARQA | 0.59 (0.53, 0.65) | 0.50 (0.46, 0.54) | $t(58.00) = 13.72, p < .001$ |
| | CRQA | 0.59 (0.53, 0.65) | 0.50 (0.47, 0.54) | $t(28.99) = 17.41, p < .001$ |
| | MdRQA | 0.60 (0.54, 0.66) | 0.50 (0.47, 0.53) | $t(58.00) = 19.90, p < .001$ |
| | Feature Fusion | 0.59 (0.53, 0.65) | 0.50 (0.45, 0.55) | $t(29.00) = 17.16, p < .001$ |
| | Late Fusion | 0.61 (0.55, 0.68) | 0.50 (0.45, 0.56) | $t(29.00) = 16.07, p < .001$ |
| Random Forest | ARQA | 0.66 (0.59, 0.73) | 0.50 (0.43, 0.57) | $t(58.00) = 17.39, p < .001$ |
| | CRQA | 0.61 (0.54, 0.68) | 0.49 (0.41, 0.56) | $t(29.00) = 19.45, p < .001$ |
| | MdRQA | 0.64 (0.57, 0.70) | 0.48 (0.41, 0.55) | $t(29.00) = 21.48, p < .001$ |
| | Feature Fusion | 0.65 (0.59, 0.72) | 0.50 (0.43, 0.57) | $t(56.00) = 22.41, p < .001$ |
| | Late Fusion | 0.65 (0.59, 0.72) | 0.48 (0.41, 0.55) | $t(29.00) = 22.59, p < .001$ |

Table 2. Prediction Correlations for RQA types with Logistic Regression and Random Forest Classification

| **Logistic Regression Model** | | | |
|---|---|---|---|
| | **ARQA** | **CRQA** | **MdRQA** |
| **ARQA** | 1 | 0.63 | 0.68 |
| **CRQA** | | 1 | 0.84 |
| **MdRQA** | | | 1 |

| **Random Forest Model** | | | |
|---|---|---|---|
| | **ARQA** | **CRQA** | **MdRQA** |
| **ARQA** | 1 | 0.61 | 0.67 |
| **CRQA** | | 1 | 0.84 |
| **MdRQA** | | | 1 |

prediction correlations, that there are few relevant differences in the information contents of different RQA approaches for detecting successful team collaboration, though these RQA metrics may still be capturing different aspects of gaze dynamics. Furthermore, from Table 1, there appears to be no substantial benefit to combining RQA features (auROC=.65, *Feature Fusion* random forest) or fusing the individual RQA model predictions via ensemble learning (auROC=.65, *Late Fusion* random forest) compared to, for instance, the ARQA random forest model (auROC=.66).

For both the logistic regression and random forest models for each of the three RQA approaches, we computed the relative importance of each RQA metric by inspecting its feature weight or influence on the trained model predictions. Across the different RQA approaches, different metrics were shown to be most important (see Figure 5). For all models and RQA approaches, vertical entropy and diagonal entropy (blue bars) consistently ranked among the top 50% in terms of importance. For all logistic regression models, the weight of these entropy metrics positively correlated with level success, meaning that an increased level of uncertainty in gaze patterns contributes to team success in this context. All other RQA metrics had variable importances across each model, perhaps suggesting that similar information could be obtained from different combinations of these metrics, given the similarity in auROC performance. This observation is also consistent with the high degree of correspondence between each model's predictions (Table 2).

## 4 DISCUSSION

In order to quantify aspects of team eye gaze dynamics, we have proposed three RQA-based methods for studying team gaze dynamics during CPS (centroid-based ARQA, CRQA, and MdRQA). Further, we contextualized each method and its associated recurrence metrics within the framework of understanding team gaze dynamics. We have also shown that team composition (based upon average prior physics knowledge score, gender composition, and race composition) was not strongly associated with the recurrence metrics assessed in this study (**RQ1**). Thus, it appears that team eye gaze

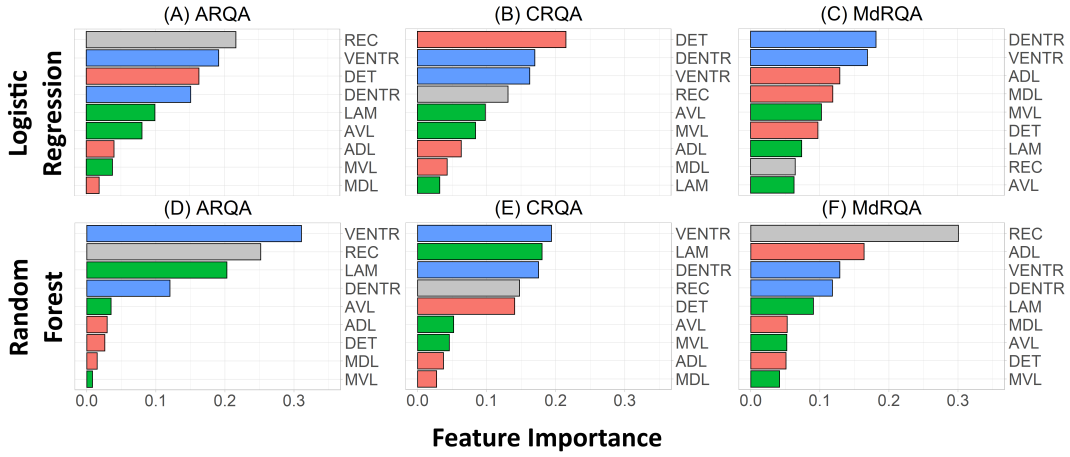


Fig. 5. Relative importance of each RQA metric in both the logistic regression (top) and random forest (bottom) models for each of three RQA approaches. RQA metrics derived from similar measures over the RQA plot (e.g., Figure 1) share a color. Recurrence (REC) is grey; Vertical entropy (VENTR) and diagonal entropy (DENTR) are blue; Determinism (DET), average diagonal line length (ADL), and maximum diagonal line length (MDL) are red; and Laminarity (LAM), average vertical line length (AVL), and maximum vertical line length (MVL) are green.

dynamics may be robust to some aspects of team diversity. Researchers such as Nielsen and Börjeson [27] argue that gender diversity and other measures of team diversity that are not primarily based on ability may be poor predictors of most outcomes. In this regard, RQA measures of team eye gaze dynamics seem to be a suitable ability-based measure of CPS which is agnostic to team diversity.

In terms of the predictive capacity of RQA-derived metrics to predict CPS success (**RQ2**), we have shown that different means of quantifying team-level dynamics using RQA have strong, yet distinct contributions to predicting team success (see Table 1). Both logistic regression and random forest classification methods show greater than chance predictive accuracy across all RQA feature sets and feature set combinations. This implies that RQA metrics of team gaze dynamics are indeed informative for predicting collaborative success, however there is still unexplained variance left to capture. The overall performances for all *original* models are minimal in terms of both raw auROC and improvement of auROC over the *shuffled* variants. Furthermore, the performance across different RQA feature sets is similar, which is surprising because of the differences in how each RQA method quantifies team dynamics. However, it is important to note that each RQA metric within both logistic regression and random forest frameworks across all RQA methods shows different patterns of variable importance and high (but not perfect) prediction correlations, indicating that each method may emphasize different yet equally important aspects of team gaze dynamics. For instance, in the context of our logistic regression model the most important feature for predicting success using centroid-based ARQA features was recurrence rate, indicating that the recurrence of gaze centroids influenced team success (Figure 5.A). However, the feature with the highest importance for CRQA was determinism, indicating that the predictability of gaze dynamics between pairs of team members was most influential to team success (Figure 5.B). Thus each RQA method yields at least some unique information for understanding team-level eye gaze dynamics.

**Limitations and Future Directions.** While RQA is a powerful tool for studying the symbolic dynamics of team members during CPS tasks, our study is not without its limitations. For instance, since the CPS data set we used was collected in a lab environment, our results may not fully translate to teams in more naturalistic environments. Another

issue is that our results are most likely dependent on the size of the grid we chose for quantifying participant screen gaze dynamics. While we believe that the 10x10 grid creates a balance between minimizing the size of each region for sensitivity and allowing regions to be large enough so that recurrence may be studied, changes in grid size may change the results of our analyses. It is also possible that individual or team-level covariates beyond prior physics knowledge, gender, and race may influence our calculated RQA metrics (e.g., experience working in a team, leadership ability, or gaming experience), none of which were included in this study. A future direction for this research would be to include more aspects of team composition (diversity and knowledge) as predictors of team dynamics and possibly perform a mediation study to understand how team composition may influence team success through team dynamics.

Another possible future direction for this research is to utilize entire recurrence matrices as predictors of team success instead of just metrics derived from each matrix. That is, each recurrence matrix plot (e.g., Figure 1) could be fed as an image into deep convolutional neural networks. In theory, this would mean that every possible feature in an RQA matrix would be used to predict team success instead of just ones defined by humans. Other researchers have used this technique for modeling EEG signals and in mechanical applications, but to our knowledge this has not been explored in the context of team collaboration dynamics [16, 18].

**Conclusion** We have shown that RQA-based methods are useful tools for studying team-level gaze dynamics during active CPS tasks involving a shared screen and that team eye gaze dynamics derived from RQA are surprisingly robust to differences in team composition. Though results demonstrate that different RQA approaches yield similar information for predicting team outcomes, we contend that certain RQA approaches are preferable to others. We argue that, while informative, aggregation of individual level dynamics prior to conducting an ARQA analysis is an overall poor choice as too much information is lost during this aggregation process, making finding meaningful relationships and interpretations difficult. Instead, we suggest that researchers aggregate pairwise metrics (CRQA) or use methods which do not aggregate at all (MdRQA) when studying team-level dynamics in a CPS context for ease of interpretation.

In this work, we argue that a holistic approach to studying multiparty interactions in CPS is essential for understanding how team-level collaboration dynamics lead to beneficial outcomes. We have shown that RQA is a useful technique for holistically studying multiparty effects in teams in the unimodal case (eye gaze dynamics), but a deeper exploration of the capacity of RQA to explain complex team dynamics in multimodal and real-world contexts is needed. We hope that the explanation of RQA and application of RQA analysis to collaborative problem solving will be helpful to other CPS researchers and helps pave a path towards more holistic multimodal and multiparty research in the future.

**Acknowledgments.** This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) (DRL 2019805) and NSF DUE 1745442/1660877. The opinions expressed are those of the authors and do not represent views of the funding agencies.

## REFERENCES

[1] Angelina Abitino, Samuel L Pugh, Candace E Peacock, and Sidney K D'Mello. 2022. Eye to Eye: Gaze Patterns Predict Remote Collaborative Problem Solving Behaviors in Triads. In *International Conference on Artificial Intelligence in Education*. Springer, 378–389.

[2] Mary Jean Amon, Hana Vrzakova, and Sidney K. D'Mello. 2019. Beyond Dyadic Coordination: Multimodal Behavioral Irregularity in Triads Predicts Facets of Collaborative Problem Solving. *Cognitive Science* 43, 10 (oct 2019), 1–22. https://doi.org/10.1111/cogs.12787

[3] Nicola C. Anderson, Walter F. Bischof, Kaitlin E. W. Laidlaw, Evan F. Risko, and Alan Kingstone. 2013. Recurrence quantification analysis of eye movements. *Behavior Research Methods* 45, 3 (sep 2013), 842–856. https://doi.org/10.3758/s13428-012-0299-5

[4] P. Arunprasad, Chitra Dey, Fedwa Jebli, Arunmozhi Manimuthu, and Zakaria El Hathat. 2022. Exploring the remote work challenges in the era of COVID-19 pandemic: review and application model. *Benchmarking* (2022). https://doi.org/10.1108/BIJ-07-2021-0421

[5] Umut Avci and Oya Aran. 2016. Predicting the Performance in Decision-Making Tasks: From Individual Cues to Group Interaction. *IEEE Transactions on Multimedia* 18, 4 (apr 2016), 643–658. https://doi.org/10.1109/TMM.2016.2521348

[6] Khaled Bachour, Frdric Kaplan, and Pierre Dillenbourg. 2010. An Interactive Table for Supporting Participation Balance in Face-to-Face Collaborative Learning. *IEEE Transactions on Learning Technologies* 3, 3 (jul 2010), 203–213. https://doi.org/10.1109/TLT.2010.18

[7] Paulo Blikstein and Marcelo Worsley. 2016. Multimodal Learning Analytics and Education Data Mining: using computational technologies to measure complex learning tasks. *Journal of Learning Analytics* 3, 2 (sep 2016), 220–238. https://doi.org/10.18608/jla.2016.32.11

[8] W. A. V. Clark and Karen L. Avery. 1976. The Effects of Data Aggregation in Statistical Analysis. *Geographical Analysis* 8, 4 (sep 1976), 428–438. https://doi.org/10.1111/j.1538-4632.1976.tb00549.x

[9] Moreno I. Coco and Rick Dale. 2014. Cross-recurrence quantification analysis of categorical and continuous time series: an R package. *Frontiers in Psychology* 5, JUN (jun 2014), 1–14. https://doi.org/10.3389/fpsyg.2014.00510

[10] Edwin S. Dalmaijer, Sebastiaan Mathôt, and Stefan Van der Stigchel. 2014. PyGaze: an open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior research methods* 46, 4 (2014), 913–921. https://doi.org/10.3758/s13428-013-0422-2

[11] Andrew T Duchowski and Andrew T Duchowski. 2017. *Eye tracking methodology: Theory and practice*. Springer.

[12] Gregory Dyke, Iris Howley, David Adamson, Rohit Kumar, and Carolyn Penstein Rosé. 2013. Towards academically productive talk supported by conversational agents. In *Productive multivocality in the analysis of group interactions*. Springer, 459–476.

[13] Lucca Eloy, Angela E.B. Stewart, Mary Jean Amon, Caroline Reinhardt, Amanda Michaels, Chen Sun, Valerie Shute, Nicholas D. Duran, and Sidney D'Mello. 2019. Modeling Team-level Multimodal Dynamics during Multiparty Collaboration. In *2019 International Conference on Multimodal Interaction*. ACM, New York, NY, USA, 244–258. https://doi.org/10.1145/3340555.3353748

[14] Jamie C. Gorman, Polemnia G. Amazeen, and Nancy J. Cooke. 2010. Team coordination dynamics. *Nonlinear dynamics, psychology, and life sciences* 14, 3 (jul 2010), 265–89. http://www.ncbi.nlm.nih.gov/pubmed/20587302

[15] Arthur C. Graesser, Stephen M. Fiore, Samuel Greiff, Jessica Andrews-Todd, Peter W. Foltz, and Friedrich W. Hesse. 2018. Advancing the Science of Collaborative Problem Solving. *Psychological Science in the Public Interest* 19, 2 (nov 2018), 59–92. https://doi.org/10.1177/1529100618808244

[16] Chongqing Hao, Ruiqi Wang, Mengyu Li, Chao Ma, Qing Cai, and Zhongke Gao. 2021. Convolutional neural network based on recurrence plot for EEG recognition. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 31, 12 (dec 2021), 123120. https://doi.org/10.1063/5.0062242

[17] Jiangang Hao, Lei Chen, Michael Flor, Lei Liu, and Alina A. von Davier. 2017. CPS-Rater: Automated Sequential Annotation for Conversations in Collaborative Problem-Solving Activities. *ETS Research Report Series* 2017, 1 (dec 2017), 1–9. https://doi.org/10.1002/ets2.12184

[18] Hsueh, Ittangihala, Wu, Chang, and Kuo. 2019. Condition Monitor System for Rotation Machine by CNN with Recurrence Plot. *Energies* 12, 17 (aug 2019), 3221. https://doi.org/10.3390/en12173221

[19] Sarah Hurwitz, Blaine Garman-McClaine, and Kane Carlock. 2022. Special education for students with autism during the COVID-19 pandemic: "Each day brings new challenges". *Autism* 26, 4 (may 2022), 889–899. https://doi.org/10.1177/13623613211035935

[20] Kevin M. Kniffin, Jayanth Narayanan, Frederik Anseel, John Antonakis, Susan P. Ashford, Arnold B. Bakker, Peter Bamberger, Hari Bapuji, Devasheesh P. Bhave, Virginia K. Choi, Stephanie J. Creary, Evangelia Demerouti, Francis J. Flynn, Michele J. Gelfand, Lindred L. Greer, Gary Johns, Selin Kesebir, Peter G. Klein, Sun Young Lee, Hakan Ozcelik, Jennifer Louise Petriglieri, Nancy P. Rothbard, Cort W. Rudolph, Jason D. Shaw, Nina Sirola, Connie R. Wanberg, Ashley Whillans, Michael P. Wilmot, and Mark van Vugt. 2021. COVID-19 and the workplace: Implications, issues, and insights for future research and action. *American Psychologist* 76, 1 (jan 2021), 63–77. https://doi.org/10.1037/amp0000716

[21] Russell Lenth. 2020. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://CRAN.R-project.org/package=emmeans R package version 1.5.1.

[22] Giuseppe Leonardi. 2018. A Method for the computation of entropy in the Recurrence Quantification Analysis of categorical time series. *Physica A: Statistical Mechanics and its Applications* 512 (2018), 824–836. https://doi.org/10.1016/j.physa.2018.08.058

[23] Lian Lian and Zhongda Tian. 2022. A novel multivariate time series combination prediction model. *Communications in Statistics - Theory and Methods* 0, 0 (sep 2022), 1–32. https://doi.org/10.1080/03610926.2022.2124522

[24] Roberto Martinez-Maldonado, Dragan Gašević, Vanessa Echeverria, Gloria Fernandez Nieto, Zachari Swiecki, and Simon Buckingham Shum. 2021. What Do You Mean by Collaboration Analytics? A Conceptual Model. *Journal of Learning Analytics* 8, 1 (apr 2021), 126–153. https://doi.org/10.18608/jla.2021.7227

[25] N. Marwan. 2008. A historical review of recurrence plots. *The European Physical Journal Special Topics* 164, 1 (oct 2008), 3–12. https://doi.org/10.1140/epjst/e2008-00829-1 arXiv:1709.09971

[26] Gabriel Murray and Catharine Oertel. 2018. Predicting Group Performance in Task-Based Interaction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM, New York, NY, USA, 14–20. https://doi.org/10.1145/3242969.3243027

[27] Mathias Wullum Nielsen and Love Börjeson. 2019. Gender diversity in the management field: Does it matter for research outcomes? *Research Policy* 48, 7 (2019), 1617–1632. https://doi.org/10.1016/j.respol.2019.03.006

[28] Marc-Antoine Nüssli and Patrick Jermann. 2012. Effects of sharing text selections on gaze cross-recurrence and interaction quality in a pair programming task. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*. ACM Press, New York, New York, USA, 1125. https://doi.org/10.1145/2145204.2145371

[29] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[30] Johanna Pöysä-Tarhonen, Nafisa Awwal, Päivi Häkkinen, and Suzanne Otieno. 2021. Joint attention behaviour in remote collaborative problem solving: exploring different attentional levels in dyadic interaction. *Research and Practice in Technology Enhanced Learning* 16, 1 (dec 2021), 11. https://doi.org/10.1186/s41039-021-00160-0

[31] Samuel L. Pugh, Arjun Rao, Angela E.B. Stewart, and Sidney K. D'Mello. 2022. Do Speech-Based Collaboration Analytics Generalize Across Task Contexts?. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. ACM, New York, NY, USA, 208–218. https://doi.org/10.1145/3506860.3506894

[32] Verónica C. Ramenzoni, Tehran J. Davis, Michael A. Riley, Kevin Shockley, and Aimee A. Baker. 2011. Joint action in a cooperative precision task: nested processes of intrapersonal and interpersonal coordination. *Experimental Brain Research* 211, 3-4 (jun 2011), 447–457. https://doi.org/10.1007/s00221-011-2653-8

[33] Meeli Rannastu-Avalos and Leo Aleksander Siiman. 2020. Challenges for Distance Learning and Online Collaboration in the Time of COVID-19: Interviews with Science Teachers. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 12324 LNCS. 128–142. https://doi.org/10.1007/978-3-030-58157-2_9

[34] Joseph M. Reilly and Bertrand Schneider. 2019. Predicting the quality of collaborative problem solving through linguistic analysis of discourse. *EDM 2019 - Proceedings of the 12th International Conference on Educational Data Mining* Edm (2019), 149–157.

[35] P Renaud, S Chartier, JL Rouleau, and J Proulx. 2009. Gaze behavior nonlinear dynamics assessed in virtual immersion as a diagnostic index of sexual deviancy: preliminary results. *Journal of Virtual Reality and Broadcasting* 6, 3 (2009). http://aix1.uottawa.ca/$\sim$schartie/Renaud-JVRB.pdf

[36] Daniel C. Richardson and Rick Dale. 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. https://doi.org/10.1207/s15516709cog0000_29

[37] Bertrand Schneider, Kshitij Sharma, Sebastien Cuendet, Guillaume Zufferey, Pierre Dillenbourg, and Roy Pea. 2018. Leveraging mobile eye-trackers to capture joint visual attention in co-located collaborative learning groups. *International Journal of Computer-Supported Collaborative Learning* 13, 3 (sep 2018), 241–261. https://doi.org/10.1007/s11412-018-9281-2

[38] V Shute, R Almond, and S Rahimi. 2019. *Physics Playground (1.3)[Computer software]*.

[39] AJ Sinclair and Schneider B. 2021. Linguistic and Gestural Coordination: Do Learners Converge in Collaborative Dialogue?. *International Educational Data Mining Society* Edm (2021), 431–438. https://files.eric.ed.gov/fulltext/ED615547.pdf

[40] Collaborative Problem Solving. 2018. Collaborative Problem-Solving. In *Encyclopedia of Social Network Analysis and Mining*. Vol. V. Springer New York, New York, NY, 229–229. https://doi.org/10.1007/978-1-4939-7131-2_100128

[41] Robert A Sottilare, Arthur C Graesser, Xiangen Hu, and Gregory A Goodwin. 2018. *Design Recommendations for Intelligent Tutoring Systems -Volume 6 Team Tutoring*. 161–168 pages. https://www.gifttutoring.org/attachments/download/3029/DesignRecommendationsforITS_Volume6-TeamTutoring_final.pdf#page=169

[42] A. Stewart and Sidney K. D'Mello. under review. CPSCoach: The Design and Implementation of Intelligent Collaborative Problem Solving Feedback. 3 (jul under review), 203–213.

[43] Angela E.B. Stewart, Zachary A. Keirn, and Sidney K. D'Mello. 2018. Multimodal Modeling of Coordination and Coregulation Patterns in Speech Rate during Triadic Collaborative Problem Solving. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM, New York, NY, USA, 21–30. https://doi.org/10.1145/3242969.3242989

[44] Shree Krishna Subburaj, Angela E.B. Stewart, Arjun Ramesh Rao, and Sidney K. D'Mello. 2020. Multimodal, Multiparty Modeling of Collaborative Problem Solving Performance. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. ACM, New York, NY, USA, 423–432. https://doi.org/10.1145/3382507.3418877

[45] Neil Vaughan and Bogdan Gabrys. 2016. Comparing and Combining Time Series Trajectories Using Dynamic Time Warping. *Procedia Computer Science* 96, September (2016), 465–474. https://doi.org/10.1016/j.procs.2016.08.106

[46] Rakesh Veerabhadrappa, Imali T. Hettiarachchi, and Asim Bhatti. 2021. Using Recurrence Quantification Analysis to Quantify the Physiological Synchrony in Dyadic ECG Data. In *2021 IEEE International Systems Conference (SysCon)*. IEEE, 1–8. https://doi.org/10.1109/SysCon48628.2021.9447059

[47] Hana Vrzakova, Mary Jean Amon, Angela Stewart, Nicholas D. Duran, and Sidney K. D'Mello. 2020. Focused or stuck together: Multimodal Patterns Reveal Triads' Performance in Collaborative Problem Solving. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. ACM, New York, NY, USA, 295–304. https://doi.org/10.1145/3375462.3375467

[48] Hana Vrzakova, Mary Jean Amon, Angela E. B. Stewart, and Sidney K. D'Mello. 2019. Dynamics of Visual Attention in Multiparty Collaborative Problem Solving using Multidimensional Recurrence Quantification Analysis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300572

[49] Sebastian Wallot and Giuseppe Leonardi. 2018. Analyzing Multivariate Dynamics Using Cross-Recurrence Quantification Analysis (CRQA), Diagonal-Cross-Recurrence Profiles (DCRP), and Multidimensional Recurrence Quantification Analysis (MdRQA) – A Tutorial in R. *Frontiers in Psychology* 9, DEC (dec 2018), 1–21. https://doi.org/10.3389/fpsyg.2018.02232

[50] Joseph P Zbilut and Charles L. Webber. 1992. Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters A* 171, 3-4 (dec 1992), 199–203. https://doi.org/10.1016/0375-9601(92)90426-M

[51] Zhiwei Zhu, Qiang Ji, and K.P. Bennett. 2006. Nonlinear Eye Gaze Mapping Function Estimation via Support Vector Regression. In *18th International Conference on Pattern Recognition (ICPR'06)*. IEEE, 1132–1135. https://doi.org/10.1109/ICPR.2006.864