Interpersonal heart rate synchrony predicts effective group performance in a naturalistic collective decision-making task

K. M. Sharika[1*], Swarag Thaikkandi[1†,2], Nivedita[3,4†], and Michael L. Platt[5,6,7,8*]

[1]Department of Cognitive Science, Indian Institute of Technology Kanpur, Kanpur, Uttar Pradesh, India

[2]Department of Biology, Indian Institute of Science Education and Research, Pune, Maharashtra, India

[3]Department of Material Science & Engineering, Indian Institute of Technology Kanpur, Kanpur, Uttar Pradesh, India

[4]Department of Theoretical Physics, University of Oxford, UK

[5]Department of Neuroscience, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[6]Department of Psychology, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA

[7]Marketing Department, Wharton School of Business, University of Pennsylvania, Philadelphia, PA, USA

[8]Wharton Neuroscience Initiative, Wharton School of Business, University of Pennsylvania, Philadelphia, PA, USA

[*]Address correspondence to

[†]Current affiliation

Abstract

Groups often outperform individuals in problem-solving. However, failure of group members to critically evaluate ideas in a discussion risks sub-optimal outcomes – a phenomenon called "groupthink". While recent studies have found interpersonal physiological synchrony to correlate with shared attention and group cohesion, whether it can track group efficacy in a collective decision-making task with an objectively defined performance measure remains controversial. To address this gap, we collected heart rate data from 58 groups (n=271) performing a task based on the hidden profile paradigm. Using multi-dimensional recurrence quantification analysis (MdRQA) and machine learning, we found that heart rate synchrony predicted the probability of groups overriding groupthink and reaching correct consensus with more than 70% cross-validation accuracy—significantly higher than that predicted by subjective assessment of team function or baseline heart rates alone. These findings demonstrate that heart rate synchrony during a naturalistic group discussion could be a biomarker of effective collective decision-making.

- **key words:** groupthink, hidden profile, heart rate synchrony, MdRQA, psychological safety

1 Introduction

Findings across multiple research domains, including organizational psychology (Woolley *et al.,* 2010) and ecology (Sasaki *et al.,* 2013), have highlighted the potential of groups to outperform individuals in solving complex problems. This collective advantage is thought to reflect the pooling of diverse skills and experiences of group members to resolve a given challenge. When team members fail to critically evaluate the ideas introduced into discussion, however, there is a risk of sub-optimal results through so-called "groupthink" (Whyte, 2015). Prior attempts to characterize what makes some groups more effective than others have so far largely relied on qualitative, survey-based assessments. An objective, quantitative account of effective collective decision making—especially in real-world contexts—is therefore an important priority for research in psychology, management, and leadership.

Over the years, multiple studies have identified physiological synchrony—similarity or correspondence in brain activity, skin conductance, heart rate or other biological processes between individuals—as a biomarker of enhanced communication, trust, and cooperation (Levenson & Gottman, 1983; Guastello *et al.,* 2006; Richardson *et al.,* 2007; Shockley *et al.,* 2003; Schippers *et al.,* 2010; Yun *et al.,* 2012; Hasson *et al.,* 2012; Spiegelhalder *et al.,* 2014; Golland *et al.,* 2015; Mitkidis *et al.,* 2015; Pan *et al.,* 2020; Bevilacqua *et al.* 2019; Thorson *et al.,* 2021; Gordon *et al.,* 2021; Tomashin *et al.,* Tomashin *et al.*). Prior studies suggest that physiological synchrony indexes shared attention, affective states, task engagement, perceived closeness and group cohesion, but whether it predicts the effectiveness of groups at problem solving is unclear. Indeed, how the association between physiological synchrony and shared cognitive and affective states pans out during process loss in which collective performance may result in an inferior outcome due to groupthink remains unexamined (Haynes & Platt, 2022).

To address these questions, we measured heart rates in three-to-six member groups performing a collective decision-making task (Fig. 1) based on the Hidden Profile Paradigm (Stasser & Titus, 1985; Larson Jr 1997; De Wilde *et al.,* 2017) in which information necessary for solving a problem is distributed unequally among group members. Specifically, participants faced the challenge of selecting the best candidate for a university faculty position based on attributes provided to them about three potential candidates. Consideration of only shared information provided to all group members implicitly encouraged groups to choose the incorrect option. By contrast, evaluation of unique information (consisting of negative attributes for two out of the three candidates and a positive attribute for the third candidate) allowed group members to arrive at the correct consensus decision (Table 1). Thus, by collecting and analyzing heart rate data of groups while they performed a task in which accurate performance depended on critical evaluation of unique information that also contradicted the option favored by the shared information alone, we tested whether heart-rate synchrony could predict the effectiveness of groups in overriding groupthink. We hypothesized that since critical evaluation of unique information requires resolving differences of opinion among group members and reaching a consensus through empathic listening and perspective-taking, something that has been previously associated with higher interpersonal physiological synchrony (Di Mascio *et al.,* 1955; Stratford *et al.,* 2012; Chatel-Goldman *et al.,* 2014; Marci, 2006;Ritov & Richter-Levin, 2014 ; Messina *et al.,* 2013), performance accuracy in this task would be associated with enhanced

heart rate synchrony between group members. By contrast, because groupthink appears to be driven by a desire to maintain group cohesion and/or avoid conflict (Esser, 1998; Baron, 2005; Pratkanis & Turner, 2013), we predicted groups showing lower heart rate synchrony would be less likely to reach the correct consensus decision.

1.1 Overview

Discrete heart rate data (beats per minute for every 2 seconds) was collected from 271 participants (88 males, 182 females) spanning the ages 18-35 years (mean ± SEM = 22 ± 0.22) using wireless sensors (Polar H10 chest straps). Participants were grouped into 58 teams of three to six members each (9 groups with n=3, 15 groups n=4, 20 groups with n=5, 14 groups with n=6) with no two members in the same group self-reporting prior familiarity with each other. The heart rate sensor's proprietary application, Polar Club, permitted simultaneous recording from all members of each team for both baseline (sitting and not talking before the group discussion, preGD data, for 5 mins) and task-related (group discussion, GD data, for a maximum of 15 mins) epochs. All participants gave informed consent to participate in the experiment. The study was approved by the Institutional Review Board of the University of Pennsylvania, Philadelphia, USA.

2 Methods

2.1 Task

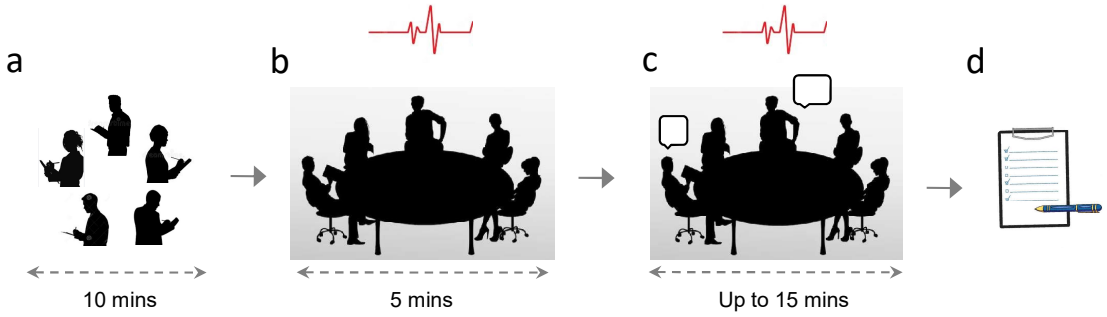


Figure 1: *Task Schematic. The group decision-making task simulated choosing the best candidate for a new faculty position at a university. a) Participants were provided 10 minutes to first read the information sheet summarizing the attributes of three potential candidates on their own, while making written notes to support ensuing discussion. Heart rate data was collected while participants were b) sitting and not talking (preGD) for 5 minutes c) involved in the group discussion (GD) which lasted for a maximum of 15 minutes across groups. d) Following the discussion, each participant had to individually respond to a set of questionnaires (GQ) on their experience of working with the group. Image used is adapted from freepik*

The group decision-making task (Fig. 1) simulated choosing the best candidate for a new faculty position at a university. Participants were provided 10 minutes to first read the information sheet summarizing the attributes of three potential candidates on their own, while making written notes to support ensuing discussion. They were explicitly told that only one candidate was most suitable for

Candidate	Information Units presented to Member 1	Information Units presented to Member 2	Information Units presented to Member 3	Information Units presented to Member 4	Information Units presented to Member 5	Information Units presented to Member 6
A	1, 2, 3, 4	1, 2, 3, 5	1, 2, 3, 6	1, 2, 3	1, 2, 3	1, 2, 3
B	1, 2, 3, 4, 5, 6, 7	1, 2, 3, 4, 5, 6, 8	1, 2, 3, 4, 5, 6, 9	1, 2, 3, 4, 5, 6	1, 2, 3, 4, 5, 6	1, 2, 3, 4, 5, 6
C	1, 2, 3, 4, 5, 6, 7	1, 2, 3, 4, 5, 6, 8	1, 2, 3, 4, 5, 6, 9	1, 2, 3, 4, 5, 6	1, 2, 3, 4, 5, 6	1, 2, 3, 4, 5, 6

Table 1: Distribution of information units (1-9) about three potential faculty candidates (A, B and C) between members (1-6) of a group in the hidden profile task. Information units denote candidate attributes that were shared between all members (in black) or unique (in green - positive attributes or in red - negative attributes) to only some individual members. Candidate A with no negative attributes was designed to be the correct answer to this selection problem.

the position and that when considering different attributes between candidates, they should value each attribute equally, even though these attributes might be qualitatively different from each other. Unknown to participants, information about the candidates was distributed unequally among group members. Some information units were presented only to some team members (unique information units) while other information units were distributed to all team members (shared information units).

Table 1 describes the distribution of information across team members. Since unique information was crucial for arriving at the correct answer, correct consensus decisions required all team members to introduce all the information available to them into discussion, and to critically evaluate each candidate's positive and negative attributes, before making a final selection. Groups were permitted a maximum of 15 minutes to discuss and render a decision, while their deliberations were recorded for both audio and video. After group discussion, each participant answered the following questionnaires on their experience of working with their group: 1) the Edmondson 7-item survey on group psychological safety (A. Edmondson, 1999; A. C. Edmondson *et al.,* 2004); 2) the three scales of the Team Diagnostic survey (Wageman *et al.,* 2005): a) process criteria for team effectiveness b) team interpersonal processes and 3) individual learning and well-being.

2.2 Data Analysis

2.2.1 Multi-dimensional Recurrence Quantification Analysis (MdRQA)

Recurrent quantification analysis (RQA) quantifies the pattern of recurrence (or how often a particular value is repeated over time) in the phase space of a dynamic system (Webber Jr & Zbilut, 1994; Marwan, Wessel, *et al.,* 2002). A phase space is a collection of all possible states of the system plotted as a function of time. How often a trajectory revisits a point in the phase space, or recurrence, indicates how different components of a multi-variate system interact and converge on the same state across time (Mitkidis *et al.,* 2015). The phase space of a multidimensional system can be reconstructed using Taken's time delayed embedding theorem (Takens, 1981), which states that if one has access to only one variable from a complex system governed by multiple interdependent variables, then one can reconstruct the dynamics of this system by utilizing (D number of) time

delayed versions of the observable x as (D-dimensional) coordinates of the phase space (Huffaker *et al.,* 2017). Let the time series be:

$$\boldsymbol{x} = (x_1, x_2, x_3, ..., x_n) \tag{1}$$

which is sampled in regular intervals in time. From this, D-dimensional vectors, $\boldsymbol{V_1}$, $\boldsymbol{V_2}$, and so on, can be constituted by estimating D-1 versions of $\boldsymbol{x}$ with time delay, τ, as below:

$$\boldsymbol{V_1} = (x_1, x_{1+\tau}, x_{1+2\tau}, ..., x_{1+(D-1)\tau}) \tag{2}$$

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V_1} \\ \boldsymbol{V_2} \\ \vdots \\ \boldsymbol{V_{n-(D-1)\tau}} \end{bmatrix} = \begin{bmatrix} x_1, x_{1+\tau}, x_{1+2\tau}, ..., x_{1+(D-1)\tau} \\ x_2, x_{2+\tau}, x_{2+2\tau}, ..., x_{2+(D-1)\tau} \\ x_3, x_{3+\tau}, x_{3+2\tau}, ..., x_{3+(D-1)\tau} \\ \vdots \\ x_{n-(D-1)\tau}, x_{n-(D-2)\tau}, x_{n-(D-3)\tau}, ..., x_n \end{bmatrix} \tag{3}$$

In the above expression, while each row represents a point in the D-dimensional phase space, each column represents the evolution of the trajectory in one of the D dimensions. MdRQA is a multidimensional extension of RQA which takes the time series of each of the group members as independent dimensions and quantifies recurrence at the group level (Wallot, Roepstorff, *et al.,* 2016; Gordon *et al.,* 2021; Tomashin *et al.,* 2022).

Measure	**Description**
Recurrence Rate (RR)	Measures density of recurrence points in the recurrence plot, indicating how probable recurrence of states is in the system.
Determinism (DET)	Measures what fraction of the diagonal line lengths are above a minimum, given, line length. $p(l)$ is the probability of a line length l. Since diagonal lines are markers of consecutive periods of recurrence in the data, determinism corresponds to the predictability of the dynamical system.
Laminarity (LAM)	Mathematically equivalent to determinism but defined for vertical (or horizontal) line lengths. Since vertical (or horizontal) lines are markers of states that do not change or change very slowly, laminarity quantifies the extent of the dynamical system being trapped in any given state for some time (Marwan, Romano, *et al.,* 2007).

Average diagonal line length (L)	Average value of diagonal line length distribution, quantifying how far in time the dynamical system is predictable.
Average vertical line length/ Trapping time (TT)	Average value of the vertical line length distribution.
Maximum diagonal line length	Maximum value from the diagonal line distribution
Maximum vertical line length	Maximum value from the vertical line distribution
Diagonal and Vertical Entropy (ENTR)	Quantifies the degree of uncertainty in the possible states and hence, the complexity of the dynamical system, using the distribution of diagonal and vertical line lengths present in the plot, respectively.

Table 2: Measures in Recurrence Analysis

Heart rate data (beats per minute) was z-transformed for the MdRQA analysis (Gordon *et al.,* 2021) since we were interested in how it changed with respect to members in a group over time rather than the absolute values. To generate the recurrence plots, time delay & embedding dimension were computed for each group separately. Time delay (tau, τ) was estimated by serially sampling values from 1 to 20 seconds and computing the first minima (first local minima or global minima when the local minima did not exist) of the mutual information between the time series of a group and a time delayed version of it (Bevilacqua *et al.,* 2019). This approach ensures that the time delayed signals are not too similar and permits the multidimensional topology of the trajectories in the phase space to unfold completely (see Kantz & Schreiber(2004), section 9.2, page 150 and supplementary for more details). For our data, tau ranged from 1 to 9 across groups. The number of embedding dimensions, m, required to adequately reconstruct the phase space was estimated using the false nearest neighbor approach (Kennel *et al.,* 1992; Hegger & Kantz, 1999; see Kantz & Schreiber(2004), section 3.3.1, page 37, figure 3.3). For our data, m ranged from 3 to 13 across groups. We chose the threshold radius, epsilon, which decides how close two points in the phase space need to be to be considered recurrent, to be such that the recurrence rate (percentage of recurrent points/ black dots in the recurrence plot) was constant across groups (10%). This allowed other variables derived from the distribution of black dots in the recurrence plot (percentage of diagonal lines (determinism) and percentage of vertical lines (laminarity), average diagonal & average vertical line, maximum diagonal & maximum vertical line, diagonal and vertical entropy; see Table 2 for their respective definitions) to be directly comparable across groups. For our data, epsilon ranged from 1.7 to 7.7 across groups. Figure 2 illustrates example recurrence plots for two groups: one that reached a correct consensus

(left panel) and one that did not (right panel).

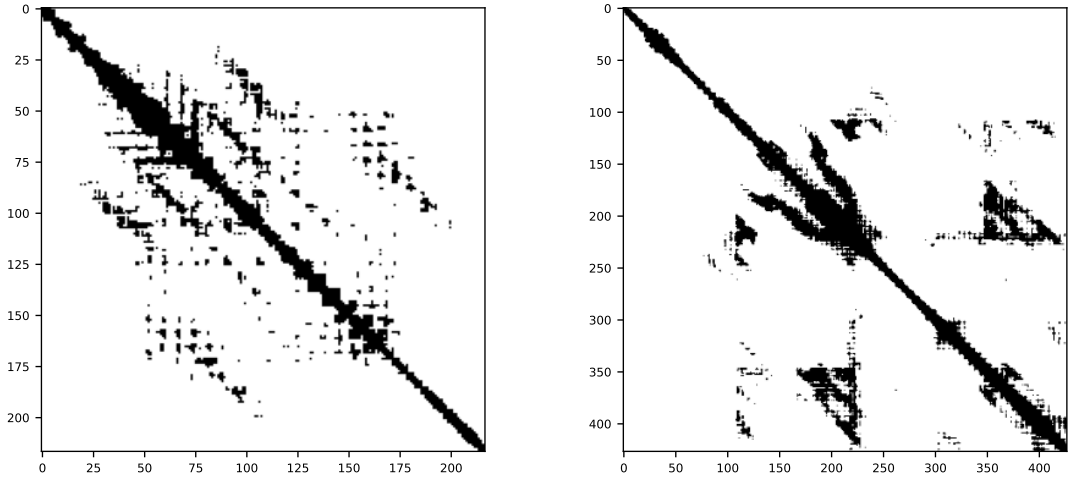
2.2.2 Sliding window MdRQA followed by Nested Cross-validation of KNN classification

Since data collection was discontinued when groups reached a consensus, the durations of group discussion varied across groups in our data set (mean ± SEM: 8.6 mins ± 36s) and resulted in recurrence plots (RPs) of different lengths. Unequal sized RPs cannot be compared directly using linear scaling because individual RPs may be sensitive to local dynamics to different extents. For the same reason, while shorter RPs may capture local dynamics, the latter may get averaged out in longer RPs. Therefore, to compare recurrent plots of different sizes, we used a sliding window approach followed by nested cross-validation of a K-Nearest Neighbour (KNN) classification procedure (k=5) that has together been shown to classify binary states of a non-linear dynamical system across varying levels of noise with high cross-validation accuracy (Thaikkandi & Sharika, 2023). A non-parametric, distance-based method of classification like KNN was used instead of using a regression model because we were interested in broadly testing if and how MdRQA variables (that are often correlated with each other for discrete heart rate data (Gordon *et al.,* 2021) predict group decision outcomes rather than studying how each variable accounts for the variability in decision outcomes individually. The sliding window approach involves partitioning a single large RP into smaller, same sized, overlapping windows along the RP diagonal and then computing the RQA variables for each of these sliding windows. The summary statistics of the different RQA variable distributions computed across the sliding windows of a given RP are then used to functionally differentiate the dynamics of a set of such RPs using a nested cross-validation procedure. Nested cross validation was used to prevent the validation set from being exposed during the training stage by keeping the training set used for feature selection separate from that used as validation set for quantifying classifier performance. This was achieved by having an inner cross-validation loop to select features via best subset selection and an outer cross-validation loop to report classifier cross validation accuracy on the validation set corresponding to that outer loop.

For each iteration of the outer loop, data was divided into three parts. While two-thirds of it was used as a training set, one-third was kept as held out validation set. Best subset selection was carried out on the training set and an aggregate performance score computed for the selected combination of features using a 2-fold repeated stratified cross validation procedure. This inner loop feature selection procedure was run for 20 iterations (resulting in 40 accuracy and ROC-AUC scores) and the combination of features having the best performance score was then chosen and evaluated on the original held out validation set of the outer loop to yield the classifier performance accuracy following a 3-fold repeated stratified cross validation procedure. The outer loop was run over 200 iterations (600 iterations of 3-fold cross-validation) to construct a distribution of performance scores (cross-validation accuracy and ROC) which was plotted as box plots to assess the ability of the classifier in discriminating effective group outcomes (i.e., correct consensus or not).

Groups with numbers of heart rate samples less than the required sliding window size for statistically significant estimates of MdRQA variables (see Supplementary section S3 for more details)

were removed from further analysis (n=2, corresponding to group discussion durations of less than 2.5 mins). We chose a common minimum sliding window size (size = 68) across all groups to first compute the MdRQA variables for each sliding window, resulting in a distribution of these variables across sliding windows for a given RP. We then used the mode of the MdRQA variable distributions for further analysis since it has been independently shown to be more robust to noise levels in the signal while discriminating between the states of a dynamical system with cross-validation accuracy higher than chance (Thaikkandi & Sharika, 2023). Please see Supplementary Information (Fig. S13, S14) for results comparing classifier performance using other measures of central tendency.



(a) (b)

Figure 2: *Recurrence plots for two groups: one that reached correct consensus (a) and one that did not (b)*

3 Results

3.1 Heart rate synchrony predicts collective decisions

Heart rate data from 12 groups could not be analyzed for measures of group synchrony due to signal disruption during data acquisition for at least one group member. Overall, data from 211 participants belonging to 46 groups (7 groups with n=3, 14 groups n=4, 16 groups with n=5, 9 groups with n=6) were analyzed using custom Python implementation of Multi-dimensional Recurrence Quantification Analysis (MdRQA) (Thaikkandi & Sharika, 2023; Wallot, Roepstorff, *et al.,* 2016). The choice of MdRQA for computing a measure of synchrony was based on previous reports demonstrating the ability of non-linear methods to better account for complexity in a phenomena (say, subtle shifts in the pattern of the data owing to its sensitivity to multiple factors) when compared to traditional linear methods (Young & Benton, 2015; Perkiömäki,2011; see supplementary section S2 for details on tests done to check whether our data satisfied a sufficient condition of non-linearity).

Based on the sliding window approach described in Methods (see Supplementary sections S2, S3, S4 & S5 for more details), only data from groups whose discussion durations lasted more than 2.5 mins were included for further analysis (n = 44). Twenty-three of 44 groups made the correct consensus decision, and 21 did not. There was no statistical difference between the group discussion durations of either set of groups (GD duration of groups that reached the correct consensus answer, mean ± SEM: 9.4 mins ± 46s; GD duration of groups that did not reach the correct consensus answer, mean ± SEM: 7.8 mins ± 55s, two sample t-test t-statistic value = -1.36, p = 0.18).

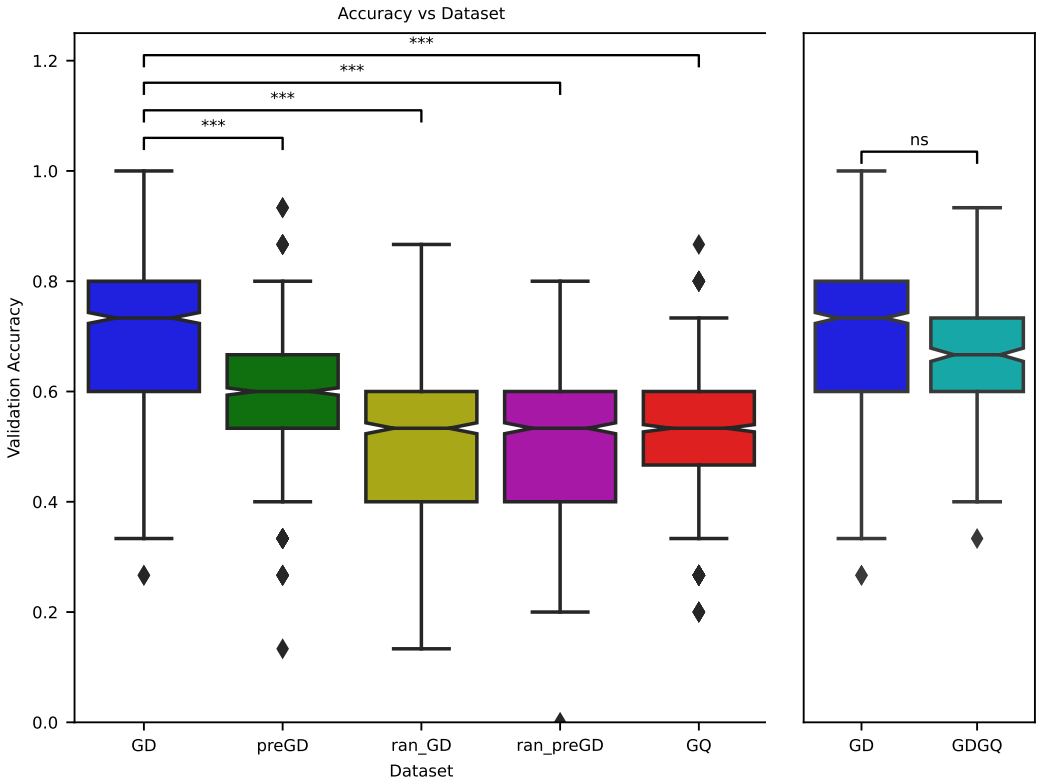


Figure 3: *Classifier performance, represented by the cross-validation accuracy for discriminating between groups that reached a correct consensus vs. not, was compared when it was trained using the following datasets: (a) MdRQA measures from time series data during the group discussion (GD) (b) MdRQA measures from time series data during baseline (preGD) (c) MdRQA measures from randomized time series data during group discussion (ran-GD) (d) MdRQA measures from randomized time series data during baseline (ran-preGD) (e) average group response to a triad of questionnaires on subjective group experience post-discussion (GQ) (f) MdRQA measures from time series data during the group discussion as well as average group questionnaire response post-discussion (GDGQ) . Same participants were sampled from each dataset in the training and validation set for creating a distribution of accuracy. Wilcoxon signed-rank test was then used for comparing the performance measure. MdRQA measures refer to the mode of the MdRQA variable distribution across sliding windows of a group RP*

The mode of each MdRQA variable distribution across sliding windows for a given RP was z-transformed before running them through the nested cross-validated KNN classification procedure

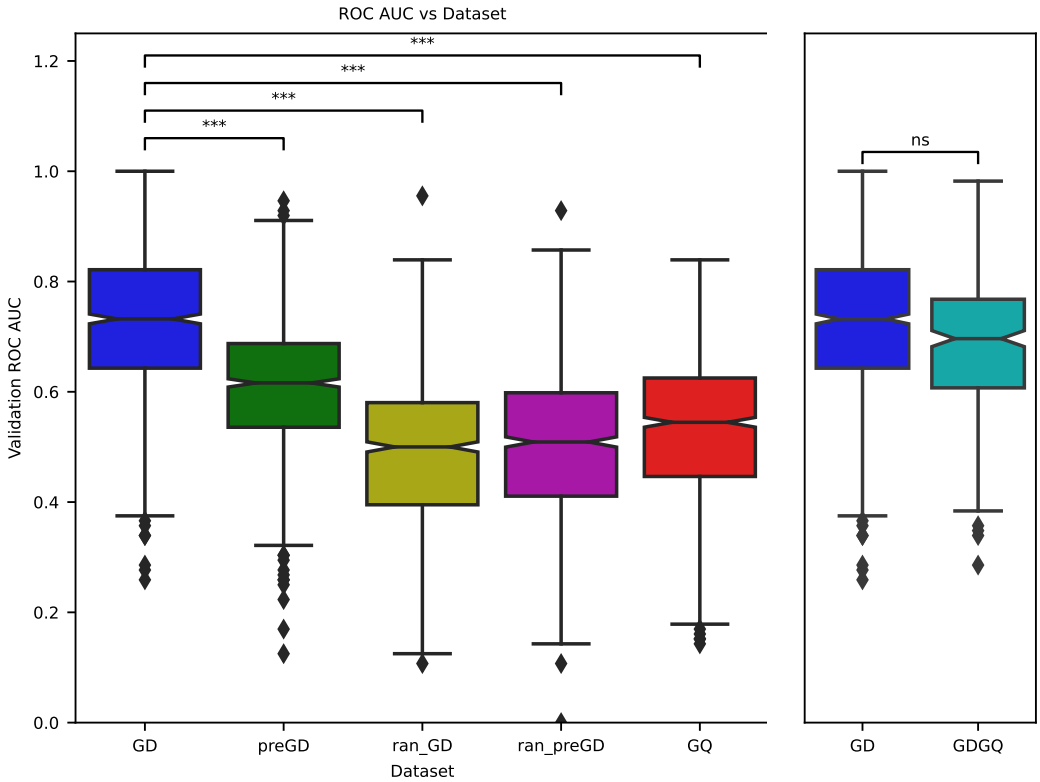


Figure 4: *Classifier performance, represented by the area under the curve of the Receiver-Operating Characteristic Curve (ROC AUC) discriminating between groups that reached a correct consensus vs. not, was compared when it was trained using the following datasets: (a) MdRQA measures from time series data during the group discussion (GD) (b) MdRQA measures from time series data during baseline (preGD) (c) MdRQA measures from randomized time series data during group discussion (ran-GD) (d) MdRQA measures from randomized time series data during baseline (ran-preGD) (e) average group response to a triad of questionnaires on subjective group experience post-discussion (GQ) (f) MdRQA measures from time series data during the group discussion as well as average group questionnaire response post-discussion (GDGQ) . Same participants were sampled from each dataset in the training and validation set for creating a distribution of accuracy. Wilcoxon signed-rank test was then used for comparing the performance measure. MdRQA measures refer to the mode of the MdRQA variable distribution across sliding windows of a group RP*

predicting whether a given group had reached correct consensus or not. Z-transformation was done so that the feature selection process was unbiased by differences in the absolute values of these variables. Classifier performance - cross-validation accuracy and discrimination ability (ROC AUC scores) - was compared for results obtained using time series data during the group discussion (GD) vs. different control conditions (described below) using Wilcoxon Signed rank test while ensuring that the groups being compared at each training and validation stage were the same across conditions (Demšar, 2006, Cohen, 1988).

We asked if heart rate dynamics between members in each group at rest (sitting and not talking)

—an index of baseline affective arousal—was sufficient to predict collective decision outcomes or whether temporal dynamics of heart rate data during group discussion was necessary. To answer this question, we ran the nested cross-validation protocol (described in Methods) separately for the following: (a) baseline (preGD) heart rate data (b) randomly shuffled heart rate data during group discussion (rand-GD) and (c) randomly shuffled heart rate data during baseline (rand-preGD). The shuffled versions were generated via random permutations to disrupt the temporal structure inherent in each times series. Comparing model performance across conditions, we found (Fig. 3) that classification accuracy was significantly higher when using MdRQA measures from group discussion compared to the control datasets: GD (median accuracy = 0.733) vs. preGD (median accuracy = 0.6) (Wilcoxon signed rank test (one sided) with Bonferroni corrected alpha = 0.0125, p = 1.923e-45, W=1.172e+05, z = 6.357, Cohen's d = 0.26), vs. rand-GD (median accuracy = 0.533) (Wilcoxon signed rank test (one sided) with Bonferroni corrected alpha = 0.0125, p = 2.398e-78, W=1.454e+05, z = 13.002, Cohen's d = 0.531), vs. rand-preGD (median accuracy = 0.533) (Wilcoxon signed rank test (one sided) with Bonferroni corrected alpha = 0.0125, p = 1.154e-78, W=1.442.0+05, z = 12.712, Cohen's d = 0.519). The set of features that got selected most frequently during feature extraction (in the inner loop during nested cross-validation procedure, see Methods for more details) for GD were recurrence rate, percentage laminarity and average vertical line length (Fig. S3). Similarly, ROC AUC values (Fig. 4) showed significantly better classifier performance in case of GD (median ROC AUC = 0.732) vs. preGD (median ROC AUC = 0.616) (Wilcoxon signed rank test (one sided) with Bonferroni corrected alpha = 0.0125, p = 6.99e-43, W=1.464e+05, z = 13.258, Cohen's d = 0.541), or rand-GD (median ROC AUC = 0.5) (Wilcoxon signed rank test(one sided) with Bonferroni corrected alpha = 0.0125, p = 3.507e-84, W=1.7e+05, z = 18.811, Cohen's d = 0.768), or rand-preGD (median ROC AUC = 0.509) (Wilcoxon signed rank test(one sided) with Bonferroni corrected alpha = 0.0125, p = 4.750e-81, W=1.685e+05, z = 18.441, Cohen's d = 0.753).

To test how the MdRQA based heart rate synchrony measures predict effective group performance when compared to the average subjective reports of group members to questions related to the experience of working with their respective groups, we ran the KNN classification procedure with nested cross-validation using group questionnaire scores (GQ) alone as well as with MdRQA measures during group discussion (GD + GQ). We found (Fig. 3) that classification accuracy was significantly higher when using MdRQA based heart rate synchrony measures during group discussion (GD (median accuracy = 0.733) vs. when using post-discussion questionnaires scores alone (GQ (median accuracy = 0.533), Wilcoxon signed rank test (one sided) with Bonferroni corrected alpha = 0.0125, p = 9.053e-11, W=1.685e+05, z = 5.222, Cohen's d = 0.213). Similarly, ROC AUC values (Fig. 4) showed significantly better classifier performance in case of GD (median ROC AUC = 0.732) vs. GQ (median ROC AUC = 0.545) (GD vs GQ, Wilcoxon signed rank test (two sided) with Bonferroni corrected alpha = 0.0125, p = 4.139e-06, W=1.0e+05, z = 2.530, Cohen's d = 0.103). However, classification accuracy achieved using MdRQA variables from group discussion (GD (median accuracy = 0.733)) did not differ significantly when compared to that including variables from group questionnaires as well (GD + GQ (median accuracy = 0.667), Wilcoxon signed rank test (two sided) with Bonferroni corrected alpha = 0.05, p = 0.727, W=7.297e+03). The set of features that got selected most frequently during feature extraction for GD+GQ were recurrence

rate, percentage laminarity, average diagonal line length and questionnaire scores from surveys on psychological safety as well as learning and well being (Fig. S7). Similarly, classifier sensitivity did not significantly differ between GD+GQ (median ROC AUC = 0.696) vs. GD (median ROC AUC = 0.732) (AUC: GD + GQ, Wilcoxon signed rank test (two sided) with Bonferroni corrected alpha = 0.05, p = 0.153, W=9.722e+03). These results suggest that while heart rate synchrony measures were significantly better at predicting group outcomes than subjective reports alone, including the latter did not help the classifier predict any better than that using heart rate data alone.

4 Discussion

In this study, we used heart rate data collected during a naturalistic, free-flowing group-discussion task, with minimal experimental intrusion or entrainment, to ask whether interpersonal heart rate synchrony between members of a group captured the dynamics of effective teams. Given that non-linear methods have been reported to account for complex phenomena better than traditional linear ones (Young & Benton2015; Perkiömäki, 2011), heart rate synchrony was computed using the multi-dimensional version of recurrence quantification analysis (MdRQA) which captures the changing dynamics of multiple team members simultaneously as opposed to relying on an averaged or pair-wise measure of association (Wallot, Roepstorff, *et al.,* 2016; Thaikkandi & Sharika, 2023). Using a group discussion task based on the hidden profile paradigm, where success was dependent on groups critically evaluating unique information against the (incorrect) option favored by shared information, we found that heart rate synchrony during group discussion, compared to pre-discussion baseline or subjective reports alone, discriminated effective groups from ineffective ones with about 70% cross-validation accuracy.

Prior work examining the role of inter-brain synchrony in group dynamics had reported its ability in predicting the overall performance of groups in a series of tasks requiring executive functioning (Reinero *et al.,* 2021). However, the inter-brain synchrony predictions were inconsistent at an individual task level. To the best of our knowledge, ours is the first study to demonstrate that in small groups of 3-6 individuals, heart rate synchrony predicts an objectively defined index of effective group performance in a task that requires groups to essentially override groupthink to reach a correct consensus outcome.

Hidden profile tasks have been used in the past to understand how information transfer takes place in small groups. In a meta-analysis of 65 hidden profile studies, Lu et al. (Lu *et al.,* 2012) showed that groups are eight times less likely to reach the optimal decision when performing a hidden profile task versus a manifest profile task, where all information is shared between all team members. Typical behavioral biases include discussing shared information earlier and more often than unique information (Stasser & Titus, 1985; Larson Jr, 1997). This pattern of behavior is found to be sensitive to social status and group hierarchy. Specifically, lower status individuals are less likely to repeat unique information than higher status individuals, suggesting that introducing unique or dissenting information may risk social costs to individuals including negative judgement, rejection, or even exclusion from the group (Stasser & Titus, 2003, Stasser, Stewart, *et al.,* 1995; Stasser, Vaughan, *et al.,* 2000; Larson Jr, 1997; 1998, 1998; Wittenbaum, 1998; Stasser & Titus,

2003; Sohrab *et al.* 2015). In this regard, success in a hidden profile task draws upon attributes previously reported to characterize effective teams, in particular psychological safety. Psychological safety describes the ability of team members to take interpersonal risks in a group by expressing, say, an unpopular opinion or unique perspective without fear of dismissal or negative evaluation (A. Edmondson, 1999; A. C. Edmondson *et al.,* 2004). In the context of our task, the ability of teams to critically examine available information and arrive at a decision contrary to the option favored by shared information could be expected to correlate with high group psychological safety. Indeed, in our data, while the classifier using heart rate synchrony measures during group discussion along with post-discussion questionnaire responses (GD + GQ) was found to perform as well as the one using heart rate data alone (GD), the top five most frequently selected features in the combined model included questionnaire scores from surveys on psychological safety as well as learning and well being (Fig. S7), hinting at the potential role of these group processes in supporting successful team performance.

5 Limitations

Since our final dataset had a relatively small sample size (n=44 groups), replication with a larger sample size is warranted. Increasing sample size may also support more complex statistical models such as decision trees and neural networks. Finally, continuous ECG data (as opposed to discrete beats per minute) could allow use of more sensitive physiological measures like heart rate variability (HRV) to better characterize physiological dynamics associated with effective group outcomes.

6 Conclusion

While research over the years has associated interpersonal physiological synchrony as a biomarker of multiple social cognitive states, whether physiological synchrony can predict the effective group decision-making remains unclear. We addressed this gap by examining heart rate data from 44 three-to-six member groups performing a collective decision-making task based on the hidden profile paradigm, in which information is distributed unequally among group members and must be evaluated critically to reach the correct consensus decision. Using multi-dimensional recurrence quantification analysis (MdRQA) and machine learning, we found that interpersonal heart rate synchrony predicted the likelihood of teams overriding groupthink and reaching the correct consensus decision, with more than 70 % cross-validation accuracy—significantly higher than either self-report assessment of team function or baseline heart rates prior to task onset. These findings are, to our knowledge, the first to demonstrate the use of heart rate synchrony as an objective marker of effective group interactions in naturalistic scenarios.

7 Acknowledgments

We thank Emily Hobart for her creative inputs on task design, Dr. Sebastien Wallot & Dr. Charles Webber for their thoughtful feedback on questions related to recurrence quantification analysis. We thank R01MH095894, R01MH108627, R37MH109728, R21AG073958, R01MH118203,866 R56MH122819 and R01NS123054, and the Wharton Neuroscience Initiative for funding the research. We also acknowledge the Wharton Behavioral Lab, University of Pennsylvania and the Remote Computing Facility at the Department of Cognitive Science, IIT Kanpur for their support.

References

1. Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N. & Malone, T. W. Evidence for a collective intelligence factor in the performance of human groups. *science* **330,** 686–688 (2010).
2. Sasaki, T., Granovskiy, B., Mann, R. P., Sumpter, D. J. & Pratt, S. C. Ant colonies outperform individuals when a sensory discrimination task is difficult but not when it is easy. *Proceedings of the National Academy of Sciences* **110,** 13769–13773 (2013).
3. Whyte, G. in *Wiley Encyclopedia of Management* 1–2 (John Wiley Sons, Ltd, 2015). ISBN: 9781118785317. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118785317.weom110157`. `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118785317.weom110157`.
4. Levenson, R. W. & Gottman, J. M. Marital interaction: physiological linkage and affective exchange. *Journal of personality and social psychology* **45,** 587 (1983).
5. Guastello, S. J., Pincus, D. & Gunderson, P. R. Electrodermal arousal between participants in a conversation: nonlinear dynamics and linkage effects. *Nonlinear dynamics, psychology, and life sciences* **10,** 365–399 (2006).
6. Richardson, D. C., Dale, R. & Kirkham, N. Z. The art of conversation is coordination. *Psychological science* **18,** 407–413 (2007).
7. Shockley, K., Santana, M.-V. & Fowler, C. A. Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance* **29,** 326 (2003).
8. Schippers, M. B., Roebroeck, A., Renken, R., Nanetti, L. & Keysers, C. Mapping the information flow from one brain to another during gestural communication. *Proceedings of the National Academy of Sciences* **107,** 9388–9393 (2010).
9. Yun, K., Watanabe, K. & Shimojo, S. Interpersonal body and neural synchronization as a marker of implicit social interaction. *Scientific reports* **2,** 959 (2012).
10. Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S. & Keysers, C. Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends in cognitive sciences* **16,** 114–121 (2012).
11. Spiegelhalder, K. *et al.* Interindividual synchronization of brain activity during live verbal communication. *Behavioural brain research* **258,** 75–79 (2014).
12. Golland, Y., Arzouan, Y. & Levit-Binnun, N. The mere co-presence: Synchronization of autonomic signals and emotional responses across co-present individuals not engaged in direct interaction. *PloS one* **10,** e0125804 (2015).
13. Mitkidis, P., McGraw, J. J., Roepstorff, A. & Wallot, S. Building trust: Heart rate synchrony and arousal during joint action increased by public goods game. *Physiology & behavior* **149,** 101–106 (2015).

14. Pan, Y. *et al.* Instructor-learner brain coupling discriminates between instructional approaches and predicts learning. *NeuroImage* **211,** 116657 (2020).

15. Bevilacqua, D. *et al.* Brain-to-brain synchrony and learning outcomes vary by student–teacher dynamics: Evidence from a real-world classroom electroencephalography study. *Journal of cognitive neuroscience* **31,** 401–411 (2019).

16. Thorson, K. R. *et al.* Oxytocin increases physiological linkage during group therapy for methamphetamine use disorder: a randomized clinical trial. *Scientific reports* **11,** 21004 (2021).

17. Gordon, I., Wallot, S. & Berson, Y. Group-level physiological synchrony and individual-level anxiety predict positive affective behaviors during a group decision-making task. *Psychophysiology* **58,** e13857 (2021).

18. Tomashin, A., Gordon, I. & Wallot, S. Interpersonal physiological synchrony predicts group cohesion. *Frontiers in human neuroscience* **16,** 903407 (2022).

19. Haynes, A. & Platt, M. *MIND TO MIND How leaders can operationalize synchrony to optimize team performance y* 2022.

20. Stasser, G. & Titus, W. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of personality and social psychology* **48,** 1467 (1985).

21. Larson Jr, J. R. Modeling the entry of shared and unshared information into group discussion: A review and BASIC language computer program. *Small group research* **28,** 454–479 (1997).

22. De Wilde, T. R., Ten Velden, F. S. & De Dreu, C. K. The neuropeptide oxytocin enhances information sharing and group decision making quality. *Scientific reports* **7,** 40622 (2017).

23. Di Mascio, A., Boyd, R. W., Greenblatt, M. & Solomon, H. C. The psychiatric interview: A sociophysiologic study. *Diseases of the nervous system* **16,** 4–9 (1955).

24. Stratford, T., Lal, S. & Meara, A. Neuroanalysis of therapeutic alliance in the symptomatically anxious: The physiological connection revealed between therapist and client. *American Journal of Psychotherapy* **66,** 1–21 (2012).

25. Chatel-Goldman, J., Congedo, M., Jutten, C. & Schwartz, J.-L. Touch increases autonomic coupling between romantic partners. *Frontiers in behavioral neuroscience* **8,** 95 (2014).

26. Marci, C. D. A biologically based measure of emotional engagement: Context matters. *Journal of Advertising Research* **46,** 381–387 (2006).

27. Ritov, G. & Richter-Levin, G. Water associated zero maze: a novel rat test for long term traumatic re-experiencing. *Frontiers in behavioral neuroscience* **8,** 1 (2014).

28. Messina, I. *et al.* Somatic underpinnings of perceived empathy: The importance of psychotherapy training. *Psychotherapy Research* **23,** 169–177 (2013).

29. Esser, J. K. Alive and well after 25 years: A review of groupthink research. *Organizational behavior and human decision processes* **73,** 116–141 (1998).

30. Baron, R. S. So right it's wrong: Groupthink and the ubiquitous nature of polarized group decision making. *Advances in experimental social psychology* **37,** 219–253 (2005).

31. Pratkanis, A. R. & Turner, M. E. Methods for counteracting groupthink risk: a critical appraisal. *International Journal of Risk and Contingency Management (IJRCM)* **2,** 18–38 (2013).

32. Edmondson, A. Psychological safety and learning behavior in work teams. *Administrative science quarterly* **44,** 350–383 (1999).

33. Edmondson, A. C., Kramer, R. M. & Cook, K. S. Psychological safety, trust, and learning in organizations: A group-level lens. *Trust and distrust in organizations: Dilemmas and approaches* **12,** 239–272 (2004).

34. Wageman, R., Hackman, J. R. & Lehman, E. Team diagnostic survey: Development of an instrument. *The journal of applied behavioral science* **41,** 373–398 (2005).

35. Webber Jr, C. L. & Zbilut, J. P. Dynamical assessment of physiological systems and states using recurrence plot strategies. *Journal of applied physiology* **76,** 965–973 (1994).

36. Marwan, N., Wessel, N., Meyerfeldt, U., Schirdewan, A. & Kurths, J. Recurrence-plot-based measures of complexity and their application to heart-rate-variability data. *Physical review E* **66,** 026702 (2002).

37. Takens, F. Dynamical systems and turbulence. *Warwick, 1980,* 366–381 (1981).

38. Huffaker, R., Huffaker, R. G., Bittelli, M. & Rosa, R. *Nonlinear time series analysis with R* (Oxford University Press, 2017).

39. Wallot, S., Roepstorff, A. & Mønster, D. 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,** 1835 (2016).

40. Marwan, N., Romano, M. C., Thiel, M. & Kurths, J. Recurrence plots for the analysis of complex systems. *Physics reports* **438,** 237–329 (2007).

41. Kantz, H. & Schreiber, T. *Nonlinear time series analysis* (Cambridge university press, 2004).

42. Kennel, M. B., Brown, R. & Abarbanel, H. D. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical review A* **45,** 3403 (1992).

43. Hegger, R. & Kantz, H. Improved false nearest neighbor method to detect determinism in time series data. *Physical Review E* **60,** 4970 (1999).

44. Thaikkandi, S. & Sharika, K. M. *Analyzing time series of unequal durations using Multidimensional Recurrence Quantification Analysis (MdRQA): validation and implementation using Python* 2023. arXiv: `2307.11675 [q-bio.NC]`.

45. Young, H. & Benton, D. We should be using nonlinear indices when relating heart-rate dynamics to cognition and mood. *Scientific reports* **5,** 16619 (2015).

46. Perkiömäki, J. S. Heart rate variability and non-linear dynamics in risk stratification. *Frontiers in Physiology* **2,** 81 (2011).

47. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research* **7,** 1–30 (2006).

48. Cohen, J. Statistical power analysis for the behavioral sciences New York. *NY: Academic* **54** (1988).

49. Reinero, D. A., Dikker, S. & Van Bavel, J. J. Inter-brain synchrony in teams predicts collective performance. *Social cognitive and affective neuroscience* **16,** 43–57 (2021).

50. Lu, L., Yuan, Y. C. & McLeod, P. L. Twenty-five years of hidden profiles in group decision making: A meta-analysis. *Personality and Social Psychology Review* **16,** 54–75 (2012).

51. Stasser, G. & Titus, W. Hidden profiles: A brief history. *Psychological inquiry* **14,** 304–313 (2003).

52. Stasser, G., Stewart, D. D. & Wittenbaum, G. M. Expert roles and information exchange during discussion: The importance of knowing who knows what. *Journal of experimental social psychology* **31,** 244–265 (1995).

53. Stasser, G., Vaughan, S. I. & Stewart, D. D. Pooling unshared information: The benefits of knowing how access to information is distributed among group members. *Organizational behavior and human decision processes* **82,** 102–116 (2000).

54. Larson Jr, J. R., Foster-Fishman, P. G. & Franz, T. M. Leadership style and the discussion of shared and unshared information in decision-making groups. *Personality and Social Psychology Bulletin* **24,** 482–495 (1998).

55. Wittenbaum, G. M. Information sampling in decision-making groups: The impact of members' task-relevant status. *Small Group Research* **29,** 57–84 (1998).

56. Sohrab, S. G., Waller, M. J. & Kaplan, S. Exploring the hidden-profile paradigm: A literature review and analysis. *Small Group Research* **46,** 489–535 (2015).

57. Isliker, H. & Kurths, J. A test for stationarity: finding parts in time series apt for correlation dimension estimates. *International Journal of Bifurcation and Chaos* **3,** 1573–1579 (1993).

58. Schreiber, T. & Schmitz, A. Improved surrogate data for nonlinearity tests. *Physical review letters* **77,** 635 (1996).

59. Lancaster, G., Iatsenko, D., Pidde, A., Ticcinelli, V. & Stefanovska, A. Surrogate data for hypothesis testing of physical systems. *Physics Reports* **748,** 1–60 (2018).

60. Kugiumtzis, D. Test your surrogate data before you test for nonlinearity. *Physical Review E* **60,** 2808 (1999).

61. Schreiber, T. & Schmitz, A. Discrimination power of measures for nonlinearity in a time series. *Physical Review E* **55,** 5443 (1997).

62. Marwan, N., Schinkel, S. & Kurths, J. Recurrence plots 25 years later—Gaining confidence in dynamical transitions. *Europhysics Letters* **101,** 20007 (2013).

63. Wallot, S. & Mønster, D. Calculation of average mutual information (AMI) and false-nearest neighbors (FNN) for the estimation of embedding parameters of multidimensional time series in matlab. *Frontiers in psychology* **9,** 1679 (2018).