JrnlID 11412_ArtID 9318_Proof# 1 - 16/03/2020

International Journal of Computer-Supported Collaborative Learning
https://doi.org/10.1007/s11412-020-09318-2

Unpacking the relationship between existing and new measures of physiological synchrony and collaborative learning: a mixed methods study

Bertrand Schneider, et al. ***[full author details at the end of the article]***

Received: 23 October 2019 / Accepted: 9 March 2020

Abstract

Over the last decade, there has been a renewed interest in capturing twenty-first century skills using new data collection tools. In this article, we leverage an existing dataset where electrodermal activity (EDA) was used to identify markers of productive collaboration. The data came from 42 pairs of participants ($N = 84$) who had no coding experience and were asked to program a robot to solve a variety of mazes. Because little is known on how physiological synchrony relates to collaborative learning, we explored four different measures of synchrony: Signal Matching (SM), Instantaneous Derivative Matching (IDM), Directional Agreement (DA) and Pearson's Correlation (PC). Overall, we found PC to be positively associated with learning gains ($r = 0.35$) and DA with collaboration quality ($r = 0.3$). To gain further insights into these results, we also qualitatively analyzed two groups and identified situations with high or low physiological synchrony. We observed higher synchrony values when members of a productive group reacted to an external event (e.g., following instructions, receiving a hint), oscillations when they were watching a video or interacting with each other, and lower values when they were programming and / or seem to be confused. Based on these results, we developed a new measure of collaboration using electrodermal data: we computed the number of cycles between low and high synchronization. We found this measure to be significantly correlated with collaboration quality ($r = 0.57$) and learning gains ($r = 0.47$). This measure was not significantly correlated with the measures of physiological synchrony mentioned above, suggesting that it is capturing a different construct. We compare those results with prior studies and discuss implications for measuring collaborative process through physiological sensors.

Keywords Biosensors · Collaborative learning · Physiological synchrony · Electrodermal activity · Galvanic skin response wristbands

Introduction

In educational research, there has been a renewed interest in leveraging new data streams for capturing students' learning processes that go beyond the acquisition of conceptual knowledge. With an ever-increasing ease of access to information, educational researchers are more and more interested in "21st century skills" (Dede 2010). These skills include (but are not limited to) students' curiosity, critical thinking, collaborative skills, grit, persistence or creativity. Having accurate and reliable tools for capturing these skills can pave the way for new kinds of instruction, for example by displaying levels of mastery to teachers through dashboards (Martinez Maldonado et al. 2012); by designing awareness tools for students (Buder 2011); or by adapting the learning environment in real time according to the learner's state (Wang et al. 2006). To reach this goal, educational researchers are starting to use multimodal sensors and learning analytics to richly capture students' behavior, for example through Multimodal Learning Analytics, (MMLA; Blikstein and Worsley 2016). MMLA is opening new doors for educational researchers, by allowing them to capture large amounts of process data that can be leveraged in adaptive systems.

Among the vast array of twenty-first century skills and MMLA measures available to researchers, this paper focuses on collaborative learning and electrodermal data. For students' collaboration, self-regulated, co-regulated, and socially shared regulation of learning play in important role in small groups (Hadwin et al. 2011). Prior work (e.g., Haataja et al. 2018) has found evidence that physiological synchrony can be associated with a group's ability to regulate itself. We build upon this work, and further explore how galvanic skin response relates to productive interactions. More specifically, we computed metrics of physiological synchrony and correlated them collaboration quality in dyads. To gain a better understanding of these indicators, we qualitatively analyzed a low performing and high performing group and analyzed how their levels of physiological synchrony changed during a collaborative learning activity. This analysis inspired a new measure of collaboration, where we captured the number of cycles of high / low physiological synchrony that each group experienced.

The article is structured as follows: first, we describe theories of collaborative learning and prior work that used electrodermal activity for studying collaborative processes. We describe the study where the data was collected, our preprocessing procedure and analyses. We then present our quantitative and qualitative results. We conclude with a discussion of our findings, highlight some limitations of our approach, and describe potential future work for this line of research.

Literature review

What does good collaborative learning look like?

While there are a wealth of theories of collaborative learning, we focus on Roschelle's (1992) framework of convergent conceptual change. In this framework, collaboration is seen as the process of constructing shared meanings for conversations, concepts, and experiences. This process has been extensively studied from a psycho-linguistic perspective and is referred to as grounding (Clark and Wilkes-Gibbs 1986). Building a common ground ensures that collaborators are on the same page and share a common definition of the terms used. From this perspective, grounding allows group members to anticipate and prevent misunderstandings. Educational researchers go beyond the psycho-linguistic definition of grounding, however, to focus on shared meaning making (Stahl 2007). Shared meaning making is associated with "the increased cognitive-interactional effort involved in the transition from learning to understand each other to learning to understand the meanings of the semiotic tools that constitute the mediators of interpersonal interaction" (p. 31; Baker et al. 1999). It gradually leads to the construction of new meanings and can result in conceptual change.

More concretely, educational researchers have identified mechanisms that promote grounding and shared meaning making: good collaborators are proficient at articulating and clarifying their thinking (Webb et al. 1995); they explicitly restructure their understanding of ideas to make visible what they do and do not know (Cooper 1999); they engage in elaborative processing, by building on their partner's ideas (Damon 1984); they actively co-construct ideas with peers (Webb and Palincsar 1996); they negotiate meanings and solve conflicts by providing sophisticated (counter-)arguments (Baker 2003). A pre-requisite for these behaviors is that group members participate equally to the shared meaning making process; a free-rider or sucker effect (Salomon and Globerson 1989) prevents good collaboration from emerging. Equality in participation among individuals in groups has been theorized to be critical for successful collaborative learning (i.e., Mutuality; Damon and Phelps 1989). This motivates the analysis described in this paper, where we hypothesize that productive interactions between learners is associated with higher physiological synchrony, and a free rider effect is correlated with less physiological synchrony. In the section below we describe prior work on physiological synchrony.

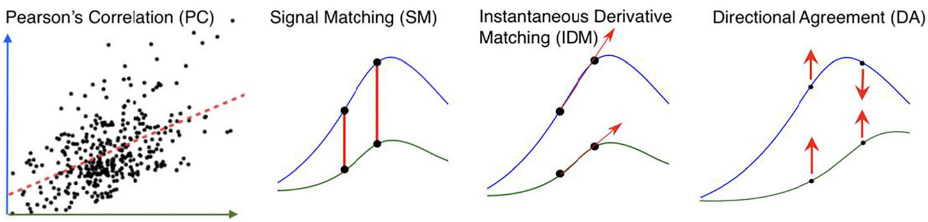
Electrodermal activity (EDA) in educational research

Electrodermal Activity (EDA) is electrical change measured at the surface of the skin, which occurs when the skin receives innervating signals from the sympathetic nervous system. The sympathetic system is activated in case of a physiological activation such as physical exertion or cognitive workload, and electrical conductance increases as the pores begin to fill with sweat. EDA is generally considered to be a reliable way of measuring sympathetic activation (Dawson et al. 2007). In educational research, EDA has been used to capture students' affective state. As an example, Arroyo et al. (2009) used data from four different sensors (camera, mouse, chair, and EDA wristband) to predict students' affects in a school setting and were able to explain 60% of the variance of their emotional state when interacting with intelligent tutors.

Of interest in the current paper is the use of physiological sensors to study the quality of social interactions. There is some theoretical basis for connecting productive collaboration with physiological synchrony. For example the "chameleon effect" suggests that partners who are in agreement tend to imitate each other (Chartrand and Bargh 1999). The "emotion contagion" effect describes situations where partners who develop empathy for each other tend to feel similar emotions (Parkinson and Simons 2009). We expect a similar effect to take place in collaborative learning settings, and physiological synchrony can serve as a proxy for capturing it.

Prior work has identified various indicators of physiological coupling indices (PCIs) and correlated these measures with different outcome measures. Pijeira-Díaz et al. (2016), for example, used Signal Matching (SM), Instantaneous Derivative Matching (IDM), Directional Agreement (DA) and Pearson's correlation Coefficient (PC).In a nutshell SM captures the difference between two EDA time-series; IDM the rate of change; DA the direction of those changes; and PC the linear relationship between them. While Fig. 1 provides a visual representation of these four PCIs, we describe these indicators in more detail under the "Methods" section.

Montague et al. (2014) found indicators of physiological synchrony to be associated with task performance for pairs of participants in a multitask environment under varied task and technology conditions. More specifically, Elkins et al. (2009)'s findings suggest that PC and



Q1 **Fig. 1** A visual representation of the four PCIs used in this article (PC, SM, IDM, DA). Blue lines represent one participant, green lines represent the other participant. Red lines represent how each PCI is computed

DA were the most useful indicators to differentiate between low and high performers. In a collaborative problem-solving task (i.e., designing a healthy, appropriate breakfast for an athlete training for a marathon), Pijeira-Díaz et al. (2016) found that IDM best predicted collaborative interactions, and DA was positively associated with learning. In a collaborative game, Järvelä et al. (2014) collected physiological data in dyads of learners and found that PC was correlated with participants' interaction and self-reported social presence. In a continuous tracking-task simulating teleoperation, Henning et al. (2001) reported that PC was a significant predictor of completion time in two-person teams. In a four-persons team, they found that PC was also associated with teamwork effectiveness during real planning meetings (Henning et al. 2009). Finally, Chanel et al. (2012) compared cooperative and competitive play and found PC to be correlated with conflicting interactions. Table 1 summarizes prior work on physiological synchrony.

In summary, there is evidence that indicators of physiological synchrony are associated with outcomes of interest to educational researchers (social interactions, learning, task performance). However, prior work has mostly looked at PC, and PC appears to be associated with a wide range of collaborative processes. Additionally, there is not a clear understanding of the difference between the four physiological indicators considered in this article (PC, DA, IDM, SM). These issues are the main focus of this article. In the next section, we present the study where the data was collected, describe our measures of physiological synchrony and correlate them with our dependent measures (i.e., task performance, learning gains, collaboration quality).

t1.1 **Table 1** Summary of Results from prior studies (reproduced and augmented from Pijeira-Díaz et al. 2016)

t1.2	Construct & Task	PCI	Study
t1.3	Team Performance (military task of building clearings for four-person teams)	DA, PC	Elkins et al. (2009)
t1.4	Team Performance (a monitoring, tracking, and resource management task)	SM, IDM, DA, PC	Montague et al. (2014)
t1.5	Collaboration, Task Performance, Learning (design of a healthy, appropriate breakfast for an athlete training for a marathon)	SM, IDM, DA, PC	Pijeira-Díaz et al. (2016)
t1.6	Teamwork (continuous tracking task simulating teleoperation)	PC	Henning et al. (2001)
t1.7	Interaction (dyads play a cooperative or competitive game)	PC	Järvelä et al. (2014)
t1.8	Completion Time (20 real planning meetings over a 6-month period)	PC	Henning et al. (2009)
t1.9	Conflicting interactions (dyads play a cooperative or competitive game)	PC	Chanel et al. (2012)

Research questions Q2

Our research questions are as follows:

- **RQ1**: Are measures of physiological synchrony (PC, DA, IDM, SM) significantly correlated with collaboration quality, task performance and learning gains?
- **RQ2a**: By looking at line graphs of physiological synchrony, can we relate events of interest to peaks (sharp increase), oscillations (jolt) and valleys (sharp decrease)?
- **RQ2b**: Do these observations vary between a high-performing and a low-performing group?
- **RQ3**: Based on RQ2a and RQ2b, can we define new measures of physiological synchrony – i.e., are cycles between low and high synchronization related to our three outcome measures?

The study

The data was collected in a prior study (for more information, see Starr et al. 2018). In this study, participants with no prior programming knowledge were paired and given 30 min to program a robot to autonomously solve a series of increasingly difficult mazes. Two different interventions were used to support collaboration (crossed in a 2 × 2 experimental design): an informational explanation on the benefits of collaboration and a visualization showing relative verbal contributions of each participant. Participants were given a pre- and post-survey on computational thinking skills and a demographic questionnaire at the end of the session. Researchers coded the quality of the collaboration, the progress of the participants, and the quality of their final code. During the study, two mobile eye-trackers captured where participants were looking, a motion sensor captured motor movement and position, and two Empatica E4 bracelets captured physiological data.

Design

The study employed a 2 × 2 between-subjects design to measure the effects of the interventions. A quarter of the dyads received neither intervention (Condition #1, "No Explanation, No Visualization"), a quarter received solely the visualization (Condition #2; "No Explanation, Kinect Visualization"), a quarter received solely the informational intervention (Condition #3; "Explanation, No Visualization"), and the final quarter received both interventions (Condition #4; Explanation, Kinect Visualization"). Participants were randomly assigned to one of the four conditions.

The Informational Collaboration Intervention consisted of the researcher verbally informing the participants about several research findings related to collaboration such as equity of speech time predicting the quality of a collaboration. Dyads not assigned to conditions with this intervention received no such information. The Visualization Intervention used audio data to display what proportion of total talk came from each participant over the past 30 s. The proportion of the screen filled with a certain color represented the relative contribution to total talk time (Fig. 2, right side).

The task required participants to use a block-based programming language to program a robot through a series of mazes (Fig. 2, middle). The robot came equipped with a microcontroller, two DC motors connected to wheels, and three proximity sensors on the front, left, and right (Fig. 2, left). Participants were first shown a tutorial video to illustrate the basic concepts of how to use a block-based programming language to program the robot. Following the video, participants had five minutes to write a simple program to move the robot past a line two feet ahead of it. Data collected during this tutorial activity is not included in our analysis.

After this initial activity, a second tutorial video describing advanced features was shown to participants and a printed reference sheet covering the material from the video was provided. The advanced features included using prewritten functions to turn and methods for checking the values of the proximity sensors. The main activity required participants to spend 30 min attempting to get their robot through the increasingly more difficult mazes. As soon as a robot could solve a maze twice in a row, the next maze was provided. Participants did not know the layout of the mazes ahead of time and were encouraged to write code that could work for any maze. During this main activity, identical hints were given at five-minute intervals to all groups. See Fig. 3 for the full procedure.

Methods

Forty-two dyads participated in the study ($N = 84$). Participants were recruited from a study pool at a university in the northeastern United States. 62% of participants self-identified as students and ages ranged from 19 to 51 years old. 60% identified as female. Participants were compensated \$20 per 90-min session of the study. No participants previously knew each other.

In addition to a variety of other sensors (see Starr et al. 2018), an Empatica E4 wrist sensor (Garbarino et al. 2014) tracked several physiological markers from each participant, including Electrodermal Activity (EDA) at 4 Hz. During the 30-min session, roughly 7200 EDA data points were generated for each participant.

Learning of computational thinking skills was assessed by a pre- and post-test consisting of four questions assessing knowledge of computer science principles such as looping, conditional statements, and interpreting code, adapted from Brennan and Resnick (2012) and Weintrop and Wilensky (2015). These questions required near-transfer and application of skills learned in the activity. Researchers evaluated the completeness of answers and how well answers demonstrated understanding of computational thinking skills. The sum of the scores was used to generate pre, post, and gains scores for each individual. Researchers double coded free response answers on 20% of the surveys and achieved an inter-rater reliability of 0.89.

While participants worked on the task, the researcher assessed their collaboration and task behaviors. The dyads' collaboration was assessed on nine scales adapted from (Meier et al. 2007)

Fig. 2 The material used in the study: the robot that participants had to program (left image), one maze (middle image) and the Kinect-based speech visualization (right side)

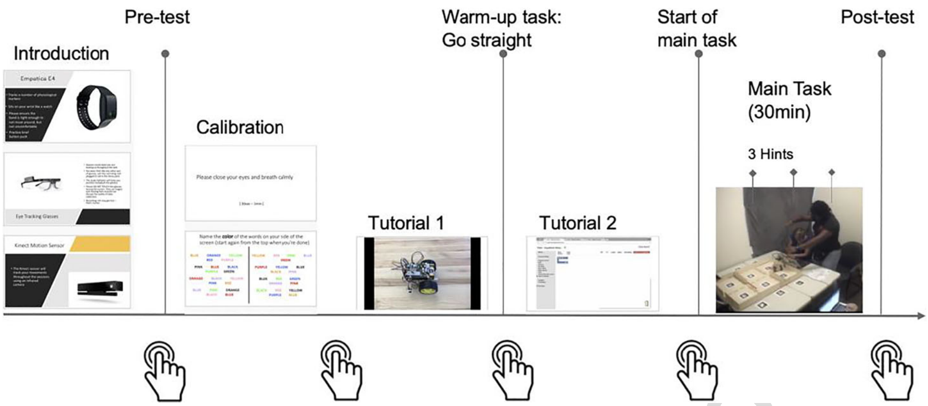


Fig. 3 The procedure of the study. Bottom row shows when participants were asked to "tag" using the Empatica wristband

on a − 2/+2 scale: sustaining mutual understanding, dialogue management, information pooling, reaching consensus, task division, time management, technical coordination, reciprocal interaction, and individual task orientation. See Meier et al. (2007) for a definition of those dimensions. Two researchers rated video recordings of the sessions using this coding scheme. The task behavior measures were task performance (number of mazes solved), task understanding (use of computational thinking concepts), and improvement over time (evidence of increased conceptual or technical understanding during the task). To calculate inter-rater reliability, a second researchers double coded 20% of the sessions from videos collected during the session and achieved an inter-rater reliability of 0.65 (Cohen's kappa), which represents a 75% agreement.

After the post-test, participants filled out a demographic survey. Following the conclusion of the activity, the final block-based code each dyad created was evaluated to determine in abstract how well the code could successfully solve different types of mazes. A rubric was created to assign a score of zero to four to evaluate the use sensor thresholds, conditional statements, looping, nesting, and generalizability. This rubric aligned with the live coding of "Task Understanding" done during the experiment, serving to ensure dyads' final products were fully evaluated. Quality of final student codes was discussed by raters until 100% agreement was reached.

Data preprocessing

We collected the following data from the Empatica wristbands: accelerometer, blood volume pulse (BVP), interbeat intervals (IBI), electrodermal activity (EDA), heart variability (HR), tag numbers to differentiate sections in each session. In this article, we focus on electrodermal activity (EDA) and describe how we preprocessed the data in the section below.

Cleaning the data

During the study, we asked participants to synchronize their sensors by pressing the button on the wristband before/after each step, which generated a tag in our dataset. By aligning these tags, we were able to synchronize the data from each participant and select one subset of the data (i.e., completing the maze task).

Before calculating our indicators of physiological synchrony, data was cleaned by removing noise. Such noise or "artifacts" can be introduced whenever an individual adjusts the sensor, knocks the wearable against something or place pressure on the device. We used EDA Explorer (Taylor et al. 2015), which is a machine learning classifier that detects noise with 95% accuracy, to remove artifacts from the data.

In the paragraph below we describe the four physiological coupling indices we explored in this article: Pearson Correlation (PC), Directional Agreement (DA), Signal Matching (SM), and Instantaneous Derivative Matching (IDM).

Computing indicators of physiological synchrony

We used PC, DA, IDM and SM because they are the most commonly measures used in the EDA literature (see Table 1). We adopted this approach because a single signal stream (i.e., EDA) may carry information from multiple phenomena, which are revealed through different analysis methods (similar to how multiple radio stations are carried over the same electromagnetic field). In this research we add to the literature by developing methods that capture different pieces of information, contained in the same EDA stream. Each physiological synchrony measure was computed using the mathematical descriptions below this paragraph.

Pearson's correlation (PC) Pearson's correlation provides an estimate of the linear relationship between two variables (here, the EDA level of both participants). For example, a positive correlation means that two participants were likely to be physiologically activated at similar times. The advantage of PC is that data from the entire session is taken into account when computing the measure. The drawback is that PC does not take time into account (i.e., the data points are looked at independently).

Directional agreement (DA) identifies whether data points from two participants increase or decrease at the same time. More specifically, each data point was subtracted with the data point that occurred right before it (Elkins et al. 2009) to determine change in signal. We then compared both individuals' data points' change. If both data points were indicated as increasing or decreasing, then this pair of points would be in "directional agreement". DA is the ratio of the total directionally agreeing pairs of points out of the total number of pairs of data points compared. The advantage of DA is that it captures whether participants EDA signals are going up or down, which makes interpretation easier. The drawback is that it doesn't take the magnitude of this change into account: dyads who experience drastic changes in their EDA get the same DA score as dyads who experience very small changes, as long as their EDA signals are increasing or decreasing at the same time.

Signal matching (SM) was used to analyze the differences in area between the curves of two participants (Elkins et al. 2009). A greater area between the curves means less synchrony while less area between the curves means higher synchrony. Thus, a negative correlation between a SM value and a qualitative measure would indicate that a small SM value means higher synchrony (and vice versa). Since individuals have different characteristics affecting their EDA signals, their signals need to be normalized to be on a comparable scale. We normalized those values using z-scores. Once the absolute differences were calculated between the data points of each individual, the overall mean difference of each pair was recorded. The advantage of SM is that it takes magnitude into account (and thus can differentiate between dyads who experience very high vs very low physiological synchrony). The drawback is that there are important individual differences in participant's electrodermal activity; this issue is somewhat alleviated by normalizing EDA scores.

Instantaneous derivative matching (IDM) calculates the similarity between the normalized slopes of two individuals (Elkins et al. 2009). The slopes are calculated as the difference between the current point and the one ahead of it. The differences between the individuals' slopes were summed up and divided by the total time range observed. The following equation was used to compute IDM:

$$\frac{1}{T}\sum_{t=0}^{T-1}|(a_{t+1}-a_t)-(b_{t+1}-b_t)|$$

The advantage of IDM is that it provides a more precise measure of the difference between two slopes (compared to DA) and it combines information about the magnitude and direction of the slopes. The drawback is the same as SM, namely that the measure is sensitive to individual differences, which means that scores have to be normalized.

In summary, the four PCIs described above capture physiological synchrony in different ways and have different (dis)advantages. Some measures focus on the magnitude of changes over time, while others focus on the direction of the signal. Because there is little to no theoretical rationale for choosing one over the others in the current literature, we analyze how these four measures relate to our dependent measures quantitatively (RQ1, RQ3) and qualitatively (RQ2a, RQ2b).

Filtering outliers

Before computing measures of physiological synchrony (DA, SM, IDM, PC), we looked for outliers. Three groups had missing EDA data and were removed from the analysis. Figure 4 (left) shows that each measure, except SM, has an outlier that was beyond two standard deviations of the mean. The right side of Fig. 4 shows the percentage of data that was removed after removing noisy artifacts (detected using EDA Explorer; Taylor et al. 2015). For our analyses, we removed outliers where too much data was either missing or noisy.

Results

(RQ1) - are measures of physiological synchrony (PC, DA, IDM, SM) significantly correlated with collaboration quality, task performance and learning gains?

We briefly summarize the main results of the study (described in detail in Starr et al. 2018). Since we are interested in the relationship between physiological data and our dependent measures, we report correlations (i.e., collaboration quality, learning gains, task performance). Our coding of collaboration was significantly positively correlated with the quality of produced code ($r = 0.52$, $p < 0.001$) as well as all three performance metrics: task performance ($r = 0.35$, $p < 0.001$), task understanding ($r = 0.53$, $p < 0.001$), and improvement over time ($r = 0.54$, $p < 0.001$). Participants gained an average of 19.8% points on the survey of

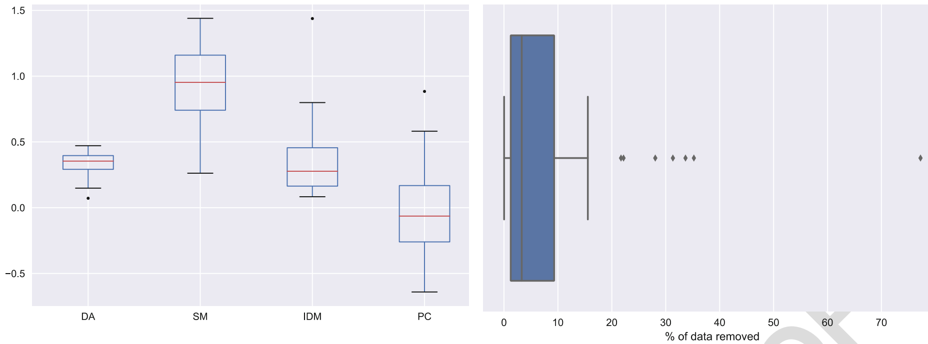


Fig. 4 Left side: boxplots for our 4 measures of physiological synchrony. We can see three outliers (one for DA, one for IDM and one for PC). Right side: percentage of data left after noisy artifacts were removed

computational thinking principles ($t = 6.18$, $p < 0.001$). The quality of the final block-based code dyads produced was significantly correlated with the number of mazes completed ($r = 0.45$, $p < 0.001$), task understanding ($r = 0.45$, $p < 0.001$), and improvement over time ($r = 0.54$, $p < 0.001$). For between conditions analyses, see Starr et al. (2018).

The first contribution of this article are the EDA analyses (not included in Starr et al. 2018). We first correlate the different PCIs with each other, and then with our dependent measures. Because all four PCIs measure physiological synchrony, we expect them to strongly correlate with each other. DA was significantly correlated with IDM: $r(35) = -0.365$, $p = 0.026$, and SM was significantly correlated with PC $r(35) = -0.658$, $p < 0.001$ (the correlations are negative, because higher DA / PC values mean more synchrony, and higher SM / IDM values mean less synchrony). There was no other significant correlation between PCIs. This suggests that different PCIs might be capturing different aspects of participants' physiological synchrony.

Correlations between PCIs and dependent measures are presented visually in Fig. 5. We found that PC was positively correlated with learning gains: $r(30) = 0.35$, $p < 0.05$; DA was positively correlated with Dialogue Management: $r(30) = 0.35$, $p = 0.063$, Reaching Consensus $r(30) = 0.36$, $p < 0.05$ and Reciprocal Interaction $r(30) = 0.470$, $p < 0.001$.

In summary, groups that were physiologically synchronized tended to achieve higher learning gains (as measured by PC) and have a better quality of collaboration (as measured by DA).

Qualitative analyses

To gain further insights into these PCIs, we chose to qualitatively compare two groups and identify which events, or behavior, seemed to be associated with more physiological synchrony. We focus on PC for these analyses, because 1) this measure was related to our main outcome measure (learning gains), and 2) PC has been found to be correlated with outcomes of interest in most prior work (see Table 1). Two groups were chosen according to the following criteria: the "best" group had to be in the top 5 groups in terms of its PC score, learning gains and collaboration quality. The "worst" group had to be in the bottom five groups in terms of the same measures. Group 35 was the best group according to these metrics and group 5 the worst. The research questions (RQs) for our qualitative analyses are as follows:

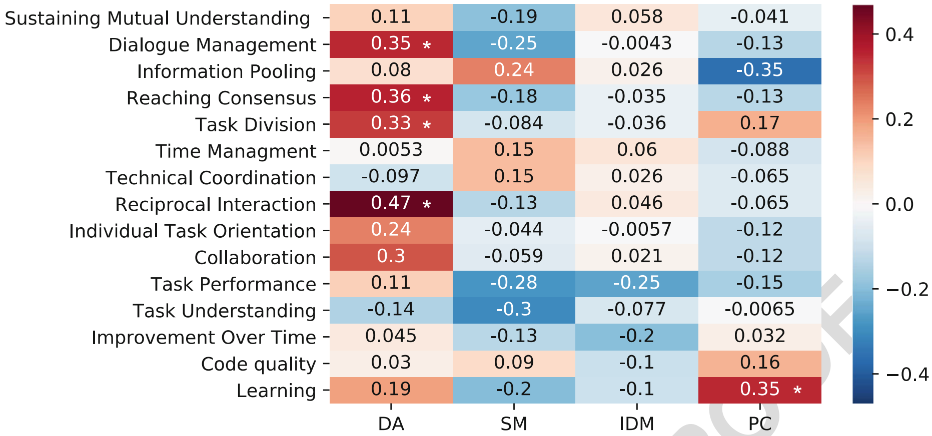


Fig. 5 Correlations between our dependent measures (collaboration, task performance and learning) and the four indicators of physiological synchrony. * $p < 0.05$, non-significant results are transparent

<u>**RQ2a**</u>: By looking at line graphs of physiological synchrony, can we relate events of interest to peaks (sharp increase), oscillations (jolt) and valleys (sharp decrease)?
<u>**RQ2b**</u>: Do these observations vary between a high-performing and a low-performing group?

Time series graphs

To explore these research questions, we created line graphs to explore synchrony values (see Fig. 6). Each data point is an averaged value of a certain amount of points determined by a rolling window. We show some graph examples below (Fig. 6). The graph on the right side has a 1 min rolling window and one on the left side a 2 min rolling window. As can be seen, a smaller rolling window means more noise. For our purposes, we wanted to find patterns between the curves of the graph and events occurring in the videos. Thus, we opted to work with the 2 min rolling windows to smooth the data and facilitate our qualitative analysis.

Video and graph observations

Selecting events of interest We first took general video notes (time, observation, subjective judgement) to understand the overall dynamic of each group. We then selected peaks, valleys and oscillations from the graphs and matched the notes via time to curves. Two researchers identified events to analyze and circled them in red for video observations. Each red circle is a section for observation in our video notes (as described below).

Video coding At every section or number labeled circle, we connect our notes from the video with the characteristics of the graph. We summarize our observations for groups 35 (high performing) and 5 (low performing) below. R refers to the participant on the right, L to the participant on the left, and F to the facilitator. In the analysis below, we underline key takeaways from the transcripts and indicate in bold the data that support these takeaways.

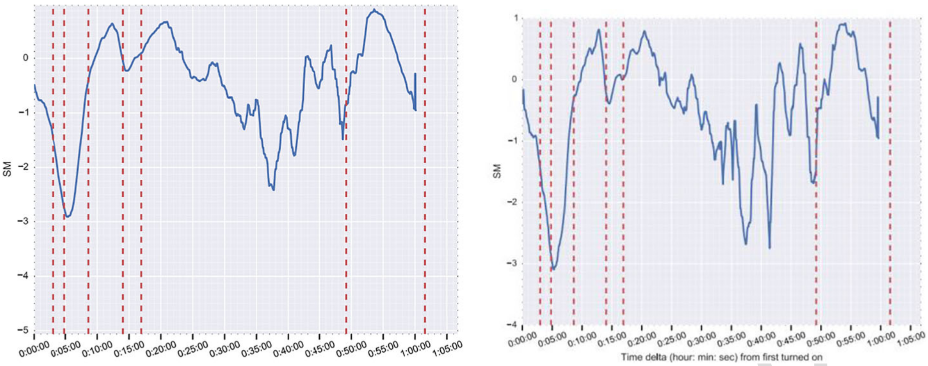


Fig. 6 Graphs with two minutes (left side) and one minute (right side) rolling windows

Group 35 (high performing group)

We qualitatively analyze the 15 events highlighted in Fig. 7 for group 35. Because the first five observations happen before the main programming task and included limited conversation between participants, we describe them more briefly. The last ten observations are described in more details because they relate to participants' collaboration and problem-solving processes. To answer RQ2a, we group these observations into three categories: peaks, oscillations and valleys. We refer to events circled in Fig. 8 using parentheses: the peak at minute 21, for example, is referred to as (7).

Peaks (events 1, 3, 5, 7, 11, 13)

In this section we analyze sharp increases in physiological synchrony. More specially, we observe the following peaks when: (1) participants are completing the baseline activities for calibrating the Empatica wristbands (meditation and Stroop tasks); (3) participants are working together to get the robot to run; they are engaged and excited to see their code work; they are working together efficiently and exhibit high levels of synchronization; (5) participants are listening to the end of the video tutorial; then the researcher introduces the main activity; participants' synchronization increases as they get ready to program the robot. Event (7) is one of the highest synchronization values for this group. Both participants are fully engaged and paying close attention to the robot's behavior.

The excerpt below shows both participants getting ready to run the code and checking with the facilitator if unconnected blocks of code on the workspace would impact the robot's behavior:

[00:18:49] L –"if we write to Gogo Board will it pick up everything on the screen? We dragged a lot of pieces
[00:18:53] out. I don't know if you can answer this.

F – "If you have spare pieces it should be fine"
R - "So it's just whatever is in the blue main box"
F – "Right"
R – "OK. So, let's run that"
[They do their first test and the robot reaches the end of the maze]

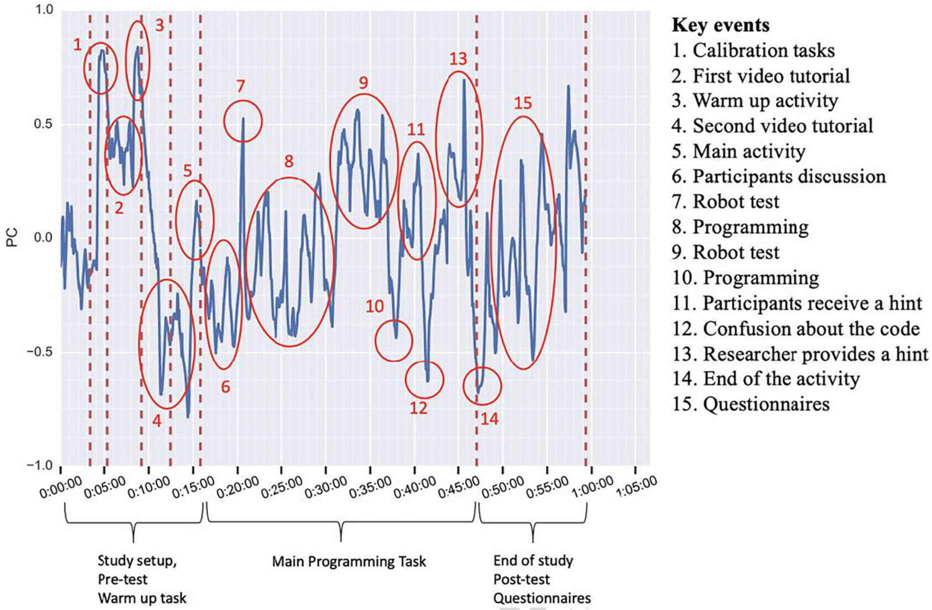


Fig. 7 Physiological synchrony of group 35 (high-performing). Selected events of interest are circled in red. Red dotted lines show when participants were asked to tag an event. Key events are provided on the right side of the graph

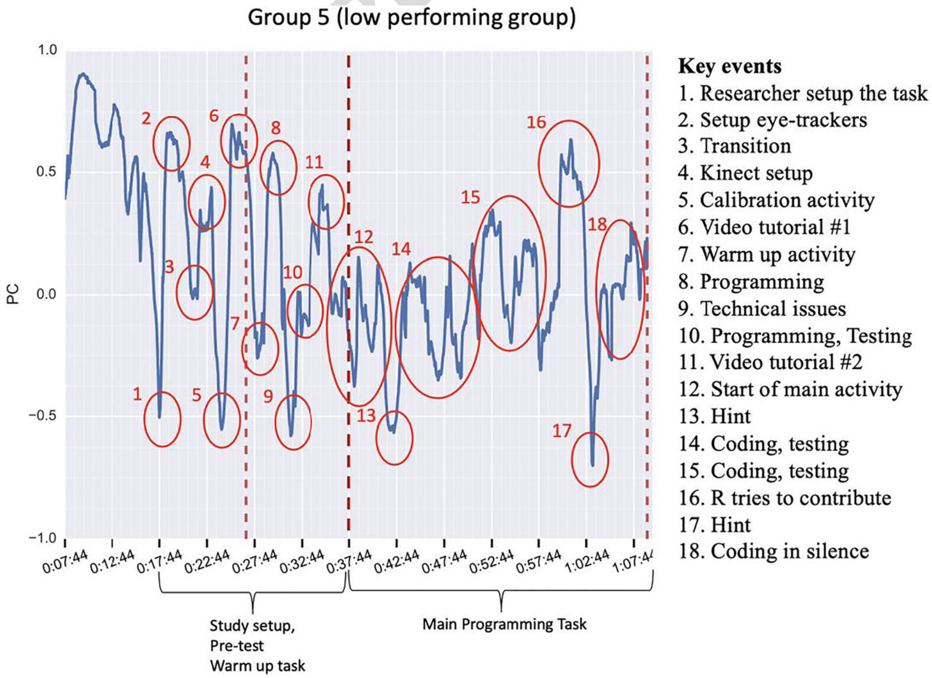


Fig. 8 Physiological synchrony of group 5 (low-performing). Selected events of interest are circled in red. Dotted lines show when participants were asked to tag an event. Key events are provided on the right side of the graph

Another peak happens at (11), when participants are transitioning from listening to the facilitator's hint and getting back to coding. The group gets on the same page regarding the hint that was provided and decide how to move forward:

[00:37:18] L - I think you should put it the whole repeat two times in the do. Is the first time it was too much of an angle.
R - OK, so let's say… then we want to turn.
L - I think you. Yeah. Perfect.
R - OK, let's. So then else, do go forward. Yeah. And then do we want this to repeat forever? Maybe that will be at the end. Let's try this. Yeah. So, are we going to write it? Let's try it again?
L - Yeah. We now have to figure out how to make go forward. But we could run this one more time.

This exchange includes several acknowledgements from both group members, which shows that they are both engaged and on the same page. The final peak happens at (13). In this event, there are two peaks in rapid succession. The first seems to be related to the facilitator telling participants that they have five minutes left to solve the current maze. The second one corresponds to participants trying to finish the final challenge:

[00:41:24] F – You have less than five minutes left. You can nest conditional statement box. Especially if else blocks to deal with more complicated scenarios. So if you have like an if else block, you could drag another if else block into that same spot
[00:41:24]-[00:45:19] R - we're just trying to figure out that logic which will probably ended up ended up being nested. So basically, we have three sensors. We're probably going to want it to mostly go forward and then turn.
L - Turn when it run into an obstacle or some sort for it. R - Yeah. So we can say.
R - It seems like the "else" is probably forward. And every time we turn we want to have this repeat. I forgot about that. Right. Every time we turn in it needs to turn twice
L - yeah. Should we try again

Again, we see participants acknowledging and building on each other's contribution. In summary, we found that peaks were associated with participants being engaged and working together towards successfully programming the robot.

Oscillations (events 2, 6, 8, 9, 15)

We observed oscillations when: (2) participants are watching a tutorial on how to program the robot. Levels of synchronization are pretty high, but they also vary, which could indicate that the two participants are paying attention to different parts of the video; (6) participants discuss strategies to solve the first maze, and they are trying to make sense of the interface. While they seem to be working on different aspects of the task and discussing different topics, they are also explaining terms to each other and clarifying what they don't know. This process of establishing a common ground (in bold, below) was associated with more oscillations:

[00:16:40]-[00:18:53] L - "Make it run for however long and then make it turn when appropriate" R - "Yeah makes sense, should we do seconds first?"
L - "Program control has if then scenario, so I think if there's a wait until, is there a time option for that?"
R - "I'll put some on the board for us to work with"
L - "Could you just click the time so I can see what is in there? Oh maybe it's that.. Wait"

R - “I imagine there’s a go forward for that”
L - “You mean go forward, in direction?”R - “No, I mean go forward for that many seconds”
L “Ohhh, got it, you probably need to attach a block to the go forward; or attach go forward to another block; go forward stuff would be in procedures if it had one like that”**R - “Ok, we should also bring the forever, right?”**
L - “oh yeah”R - “Maybe we should experiment and hitting that wall”
L - “Ok; if you have a time block now, do you want to do go forward for 2 s?”
R - “Should we edit go forward?”
L - “Sure”
R - “Because I’m not sure how to edit the time”
L - “Me neither”R - “Do you know where the expand is?”
L - “What do you mean by expand?”
R – [Explains expand]
L - “Ohh…”R - “We may not need this”

Event (8) shows a transition from testing the robot to revising the code. Participants are exploring and explaining blocks to each other. Both are engaged with the code; the participant on the right is slightly more engaged because she is controlling the mouse (as shown by the transcript below):

[00:20:36] [00:23:03] R - **if we look at the second one, we can see if there’s a time component and then go forward**. Oh good. Wonder why that’s not here. So this has turned on. There’s gotta be a way here to have a big turn on for like three seconds. Right? You can also try using sensing. Perfect. Wonder why that’s not in here. Maybe we should edit it. So I’ll bring this here. I’m going to take out the go forward.
L - I think this one is just choosing what direction it’s going and it looks like it’s reversed or if else these conditions are met then it’s the other direction, but I don’t think this is a timed one this example.
R - Isn’t that the basic go forward. **Yes. These go forward and go backwards**. I imagine would be similar. It’s just that this one has a time component and this one it doesn’t. Yeah, it doesn’t. So we edited this one to introduce the time component that could help. **Assuming we want to do it for three seconds or we might want to see if we can have it sense the wall in front and then turn here forward**. Yeah, sorry. Sorry. I’m just trying to figure out how to expand it so that we can then add this
L - Do you like get in these big blocks because you have to add it to a block where you don’t expand the block to yourself. You add the block to another block.
R - This small block is representative of this whole big block.
L - OK.

Event (9) shows participants working together and trying to implement the hint provided by the facilitator. By doing so, they are building a common understanding of how the robot work (in bold below):

[00:31:11] R – I want to figure out where the front sensor is - **Do you want to figure out where the front sensor is?**
L - Cause if we could detect the square below it and turn at that
R - Right. So is there a sensor below?
L - I’m not sure. I think now it looks like it’s just the front and the sides
R - So should we move it? Maybe we can test it here versus here, and see what it detects?
L - We’ll leave it there. I’m the sensor board up. It’s just fluctuating a lot right now.
R - oh ok. I’m going to have it go slowly. Have it go to the end because I’m not seeing a huge difference. **OK. Yeah. So that’s sensor one,**
L - So once it hits the wall it, the sensor goes up a lot

(15) The facilitator administers the remaining questionnaires.

In summary, this group seems to exhibit oscillations synchrony when they are in the process of building a common ground – they are actively building an understanding of the hardware and code in front of them. Additionally, we observed oscillations when one participant was more engaged than the other.

Valleys (events 4, 10, 12, 14)

We observe valleys (i.e., moments of low physiological synchronization) when: (4) participants were watching the second video tutorial; (10) the dyad was very involved with testing and trying to understand their situation, and then received a hint - which might have disoriented the two. The hint is followed by collaboration right afterwards – increasing level of physiological synchrony to observation (11) which has about the same synchronization values as (9).

(12) This valley represents a transition where there is a realization of what is happening with the robot - the stop button makes it turn at the wall which is what they want - but they don't understand why, and they go forward with it. The drop could be from R demonstrating what she realized as L is catching up with understanding. The rise can be from digging into this understanding.

[00:38:31]	R presses stop to show L that the robot all of a sudden turns and does what they need it to do; L says "Ohh" - could be the different levels of understanding or different approaches to understanding the situation

(14) The activity ends around 47 min, and there is a transition between last activity and the In tagging procedure; participants remove the mobile eye-trackers.

In summary, we observed the following trends in Group 35: when participants reacted to an external event (e.g., following instructions received a hint, running the robot), we observed higher synchrony values (peaks 3, 5, 7, 9, 11, 13). When they were watching a video or collaborating, we observed oscillations on the graph (2, 4, 6, 8, 9). When participants were programming and / or seem to be confused, we were more likely to observe low synchronization values (valleys 10, 12, 14). These observations are in line with what we would expect: when there is a salient event or when both participants are engaged, their levels of synchronization rise; when they work together for longer periods of time, these levels fluctuate; when there is a transition or when they seem confused, synchronization drops. In the section below, we compare these observations with a low performing group.

Group 5 (low performing group)

We identified 18 events of interest for group 5 (Fig. 8). The first 11 observations took place before the main programming task; we describe them more briefly. The last seven observations are described in more details because they relate to participants' collaboration and problem-solving processes. To answer our first research question, we group these observations into three categories: peaks, oscillations and valleys.

Peaks (events 2, 4, 6, 8, 11, 16)

The first 11 events take place before the main activity are about setting up the sensors, calibrating them, showing videos tutorials to participants and providing a warm-up task: (2), participants are asked to wear the mobile eye-trackers; (4) the researcher explains that a Kinect sensor will be tracking their body postures and gestures; (6) the researcher tells participant that they are going to watch a video tutorial on how to program the robot; (8) the researcher introduces the first activity; (11) the dyad is attempting to test the robot so both are actively involved (L is holding the wire, R is debugging and running the code). The main event (16) happens when R tries to contribute more to the conversation (in bold below), but L dismisses her suggestions:

[59:44] F - It's time for the next hint. So, I started using these, but you try using the if do - and also the if do else do blocks those will be really helpful in achieving your goal.
L - Okay.
F - Have you found that? If do else do block yet? It should be under program control or common blocks. Yeah the if do else do.
L - Okay, alright. If do, else do.
F - you have 10 min left.
L - Right. if it's less than 100. Ah okay. If we can do. No, sorry. So it's greater than 500. Do right, otherwise go forward. Like that. Let's see.
R - What?
L - Let it go. [the robot got stuck]
R - Okay.
L - Try that again. Can you try guiding it?
R - Yeah
L - Still around.
R - Maybe we can decrease the sensor.
L - No. It's not with the sensor. It's...
R - Try right? And then back?
L - Why would we want it to go back?
R - I don't know.
L - Let's try this.

In summary, we found that group 5 exhibited peaks while they were watching video tutorials, programming, or when the second member of the dyad was trying to contribute to the conversation.

Valleys (events 1, 3, 5, 7, 9, 13, 17)

In this section we analyze the valleys of group 5: (1) the researcher is setting up the tasks; participants are waiting; (3) synchronization drops as participants are waiting for the researcher to setup the remaining sensors; (5) participants are completing the baseline activities for calibrating the Empatica wristbands; (7) participants are watching a tutorial on how to program the robot. Levels of synchronization are pretty high, but they also vary, which could indicate that the two participants are paying attention to different parts of the video; (9) the dyad attempts the first activity (quiet conversation, R is looking at the handouts; L directly asks questions to the facilitator instead of talking to R); (13) participants get a hint:

[00:41:44] F - Every five minute you get a hint. So, can I see your coding page? Yeah, so, you're exactly on the right track. Your first hint is that **you need to use the sensor to help achieve the task**. Keep using a, good job. I'll give you another hint in 5 min.

At event (17), they receive another hint from the facilitator, which causes some confusion:

[63:00] F - Okay, So you have 10 min left and your last hint is that it may be useful for you to nest if do- and if do else do statements with each other. So try experimenting with that as well!
L - if do, if do else... [sounding confused]

In summary, we observed valleys primarily when participants were waiting for instructions or when they were receiving a hint. One possible interpretation is that the hint was useful to one participant, but not the other.

Oscillations (events 10, 12, 14, 15, 18)

In this section we describe oscillations: (10) because of some technical issues, the facilitator has to intervene to fix them; (12) participants are listening to the gogo board tutorial; they are waiting for the facilitator to setup the next task; they are trying to figure out which sensors is where (Fig. 9):

Fig. 9 Group 5 [00:37:44]-[00:40:44] the participant on the right asks his partner to hold the chord while they are testing the robot

[00:37:44]-[00:40:44]	L - Okay. Should I go? Okay. So, what do we want is you want to go straight go for it. When does it- when do you want it to turn right? R - To go forward... do you want to try that one? L - So the sensor is set in the front? Okay yes […] **L - Alright. Okay. So, let's see which sensor it's seeing. Would you mind holding it again?** Make sure it doesn't. Okay, so that sensor, what? Can you turn it off? Okay so let's see. Let's do it again. Ready to turn off. Yeah, should turn right. Let's see. F - Aren't you? So that's where we can give you your program. **L - There are no other sensors. Okay, so okay so, that's sensor four. Left is sensor four.**

(14) R is repetitively testing the code; L is reading the cheat sheets; there is little to no conversation between the two participants:

[00:43:44]	L - Okay. So, the sensor. Yes. So, it fits. Yeah. So, turn it right there. Okay. I think I put it wrong. Let's try it again. Yeah, you mind I'm trying it again? okay. [L is mumbling – there is no contribution from right; he's making incremental changes and testing the robot; right is mostly just holding the chord when they're testing the robot]

(15) R is in control of the mouse and does all the programming; he corrects L contribution when she tries to suggest a solution:

[50:00]	R – I'm going to try that. Do you see what's wrong? I can't figure it out. L – can you try to make it turn right **R – no it's not - what do you mean?** L – just turn right **R – if it turns right, it just turns right; look. You see? What I mean to say is that when it sees that thing it should stop. I'm trying to make it stop and turn right.** L – Oooh.

(18) participants are coding in silence. Eventually, they successfully make the robot achieves its goal. The person on the left (L), however, did all the coding – R rarely contributed. The interaction was one-sided.

To answer RQ2a ("By looking at line graphs of physiological synchrony, can we relate events of interest to peaks (sharp increase), oscillations (jolt) and valleys (sharp decrease)?"), we found that group 5 exhibited peaks when the sensors were calibrated (2,4), participants were watching video tutorials (6,11) and programming (8,16). We observed jolts during iterations of programming / testing the robot (10,14,15). Valleys were associated with transition phases (1,5,7), technical issues (9) or when the group received a hint (13,17).

To answer RQ2b ("Do these observations vary between a high-performing and a low-performing group?"), we observed that patterns of physiological synchrony were different between group 5 and 35. For example, receiving hints was an opportunity for group 35 to synchronize; for group 5, we observed sharp decreases when the group received a hint (e.g., events 13, 17). We interpret these differences as being caused by a free rider effect: in group 5, one participant wrote most of the code and declined contributions from his partner. We observed one moment where the second participant tried more clearly to contribute (event 16), which was associated with the high physiological synchrony. One trend that was similar across both groups is the tendency to exhibit a jolt when they were programming and testing the robot – most likely representing quick cycles of individual exploration followed by episodes of collaboration. Finally, another difference is that group 35 seemed to increase its physiological synchrony during the main activity – whereas group 5 had a low score throughout the activity. These two observations provided the basis for new measures of collaborative synchrony. In the next section, we examine whether increase of physiological synchronization (i.e., positive slope) and cycles of low/high synchronization (i.e., cycles of individual exploration followed by collaboration) relates to our dependent measures.

In the next section, we explore our third research questions (RQ3: "Based on RQ2a and RQ2b, can we define new measures of physiological synchrony – i.e., are cycles between low and high synchronization related to our three outcome measures?")

New measures of collaborative synchrony: Slope and cycles of PC

In this section we test two hypotheses generated through the qualitative analysis above: 1) good collaborative learning groups tend to become more and more synchronized over time; 2) good collaborative learning groups go through more cycles of (dis)synchronization compared to low performing groups. Because we focused on PC for the qualitative analysis, we also look more closely at this PCI in this section. The first hypothesis was operationalized by fitting a regression line during the main activity and correlating the slope with our dependent measures. The second hypothesis used the methodology developed by (Schneider 2019); we aggregated the data in 30s time windows and computed the number of inflection points in a time series graph (i.e., how many times did the curve went up and down during the main coding activity). This roughly represents cycles of low/high physiological synchrony. For both measures, we found the same outliers as in Fig. 4. They were removed from our analyses below.

For the first hypothesis (i.e., do slopes relate to outcome measures?), we only found one significant correlation between the slope of PC and Reaching Consensus $r(30) = -0.370$, $p = 0.037$. The negative correlation indicates that groups who better managed to reach a consensus decreased their physiological synchrony over time (which is the opposite of our prediction). No other correlation was found to be significant.

For the second hypothesis (i.e., does the number of cycles relate to our outcome measures?), we found that the number of PC cycles was significantly correlated with a number of outcome measures: Sustaining Mutual Understanding $r(30) = 0.642$, $p < 0.001$, Dialogue Management $r(30) = 0.683$, $p < 0.001$, Information Pooling $r(30) = 0.395$, $p = 0.025$, Reaching Consensus $r(30) = 0.450$, $p = 0.01$, Individual Task Orientation $r(30) = 0.351$, $p = 0.049$, overall Collaboration $r(30) = 0.570$, $p < 0.001$, Learning $r(31) = 0.466$, $p = 0.006$. This measure was not correlated with PC itself ($r = -0.161$, $p = 0.341$) or with the slope measure described above ($r = 0.010$, $p = 0.955$) – suggesting that it's capturing a different construct. Additionally, when applying this measure to the other PCIs, we found similar results (see Fig. 10):

These findings suggest that productive groups do not become more synchronized over time The number of cycles between moments of low and high physiological synchrony, however, seems to be an important feature of high-quality interactions (both in terms of participants' quality of collaboration, but also their learning gains). We discuss these results below.

Discussion

Our initial quantitative analyses suggest that PC is positively correlated with learning gains and DA is associated with our dyads' quality of collaboration. It should be noted that our

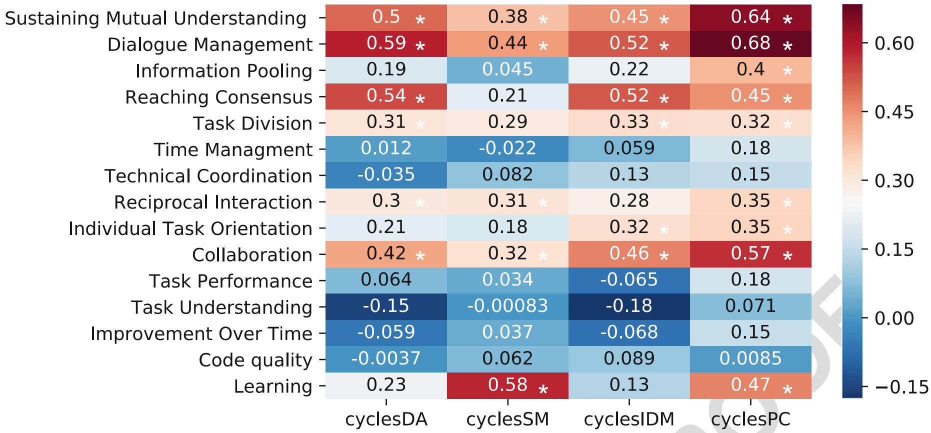


Fig. 10 Significant correlations between the number of cycles of low / high physiological synchrony for each PCI, with our dependent measures (left side). * $p < 0.05$, non-significant results are transparent

correlations did not agree with prior research. In other studies (e.g., Elkins et al. 2009; Henning et al. 2001; Pijeira-Díaz et al. 2016), group work was found to be positively correlated with SM, IDM, DA and PC. Some of these differences are likely caused by how the constructs were operationalized. Pijeira-Díaz et al. (2016), for example, used self-report scales for capturing social interactions while we applied a validated rating scheme in the learning sciences (Meier et al. 2007). Task performance and learning gains also depend on the nature of the task and can vary widely in how they are measured (e.g., completion time, success, factual knowledge, transfer questions, etc.). But it is striking to see that compared to prior work, our four measures of physiological synchronization seem to be sensitive to different outcomes measures. One interpretation is that different PCIs might be capturing different constructs, which is also supported but the fact that most PCIs were not correlated with each other. Another interpretation is that these findings are the results of imperfect measures that capture different aspects of a group's physiological synchrony. In any case, additional research is needed to further unpack the differences between these PCIs across various settings.

Our qualitative analyses further illuminated those results. We compared a high-performing group with a low performing group. Initial analyses indicate that for the high performing group, participants had higher synchrony values when reacting to external events, oscillations when they were watching a video tutorial or collaborating, and lower synchrony values when they were programming or seemed confused. These results are in line with what we could expect: levels of physiological synchrony increase when participants are working together and decrease when they working independently. We also observed that participants in the high performing group seemed to become more synchronized over time. For the low performing group, we found that low synchronization values can indicate a free rider effect. Unlike the high performing group, participants exhibited less synchronization when reacting to an external event, such as receiving a hint. This indicates that we can potentially detect poor collaboration (or a free rider effect) through levels of physiological synchrony. Interestingly, group 5 and 35 were not representative of the results found in RQ1: overall, they did not visibly differ in their average levels of physiological synchrony. This highlights a limitation of traditional statistics (i.e., reporting results on averages), and how the aggregated findings do not always apply to specific groups.

More importantly, the qualitative analyses inspired two new measures: increased synchronization over time (through the slope of a regression line) and cycles of high / low synchronization. The first measure did not yield any significant results, except with a subdimension of our coding of collaboration (reaching consensus). The correlation was negative, suggesting that groups that were better able to reach a consensus saw their physiological synchrony decrease over time. One interpretation is that they created conventions on how to work together early on (i.e., "quick consensus building"; Weinberger and Fischer 2006), which was accompanied with higher levels of synchrony, and they did not revisit these conventions later on, which was accompanied with lower levels of synchrony. Additionally, we captured the number of cycles between low and high levels of synchrony and found this indicator to be strongly correlated with outcome measures. These findings connect with previous literature showing that collaborative problem-solving is a cycle between moments of understanding and non-understanding (Miyake 1986), and that ideal cycles of communication are related to group performance (Tschan, 2002). To our knowledge, it is the first time that this relationship is established for physiological data. Q3

Limitations

In working with EDA data to understand technology-supported collaborative learning, the field is still at the stage where the reliability and validity of our analysis methods are being refined. In studies such as the present one, there are typically multiple dependent measures that are correlated with each other, for example to explore possible links between EDA synchrony metrics and collaboration and learning outcomes. In this paper, we tested the relationship between 15 dependent measures and 8 measures of physiological synchrony. Performing multiple such statistical tests of correlation increases the chance of statistical errors and reduces the reliability of the analysis. When dealing with so many variables there is increased change of Type-1 errors. Additionally, the qualitative analysis was performed on 2 of the 42 groups, and these were specifically chosen because we wanted to understand the differences between a high performing and a low performing group (in terms of their quality of collaboration and learning gains). We acknowledge that these results are not representative of the entire sample, and that we cannot draw broader conclusions from these analyses. While the findings presented in this article are promising, we recognize the limitations of our analysis, and we envision these to be used as directions for future research. As such, their generalizability should be tested through further studies.

Conclusion

In conclusion, this article replicated prior findings showing that physiological synchrony can be predictive of collaborative learning. Qualitative analyses highlighted important differences between low performing and high performing groups and suggest that we can potentially identify (un)productive collaboration through physiological data. Finally, we developed new physiological measures of collaboration and found that the number of cycles between low/high synchrony was strongly associated with collaboration quality and learning gains.

Those results are encouraging, especially in the context of developing real-time, just-in-time, personalized feedback to students. There is some preliminary evidence that high-frequency data can indeed improve collaboration: Bachour et al. (2010), for example, used a representation of microphone data to display the verbal participation of each group member; they found that displaying this data in real-time promoted more equal participation during meetings. Schneider and Pea (2013) used a dual eye-tracking setup to display the gaze of pairs of participants in real time; they found that this "gaze awareness tool" helped learners build common ground (by seeing the gaze of their partner in real time, they could more easily follow their thought process). Abrahamson et al. (n.d.) used motion sensor data to provide an embodied experience of the concept of ratio, which supported the acquisition of this concept by young learners. These three (non-exhaustive) examples suggest that sensor data has the potential to support learning of various skills and concepts. An open question is whether physiological data can provide the same benefits.

We can imagine leveraging these measures to develop dashboards for teachers and awareness tools for students (Buder 2011) – which has not been explored for physiological data. For example, being able to show these indicators in real time could be an interesting way of supporting remote or co-located collaboration (especially in contexts where facilitators are coaching participants to learn good collaborative skills, or in contexts where participants are discussing emotionally-charged topics). This unexplored area of research could potentially help students study their own collaborative behaviors and reflect on how to improve them.

In conclusion, this work opens new doors in capturing real-time indicators of collaboration and potentially using these indicators in real-time to support social interactions. These indicators could be displayed to learners and teacher to promote awareness of how collaborative processes can be co-regulated, or integrated into existing learning environments to make them more adaptative.

Acknowledgements This work was funded by the Harvard Graduate School of Education (HGSE) through the Dean Venture Funds. We also thank the Harvard Decision Science Lab (HDSL) for their support throughout this project.

References

Abrahamson, D., Trninic, D., Gutiérrez, J. F., Huth, J., & Lee, R. G. (n.d.). *Hooks and shifts: A dialectical study of mediated discovery* (pp. 1–31). Knowledge and Learning: Technology.

Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson, R. (2009). *Emotion Sensors Go To School., 200*, 17–24.

Bachour, K., Kaplan, F., & Dillenbourg, P. (2010). An interactive table for supporting participation balance in face-to-face collaborative learning. *IEEE Transactions on Learning Technologies, 3*(3), 203–213. https://doi.org/10.1109/TLT.2010.18.

Baker, M. (2003). Computer-mediated argumentative interactions for the co-elaboration of scientific notions. In J. Andriessen, M. Baker, & D. Suthers (Eds.), *Arguing to Learn: Confronting Cognitions in Computer-Supported Collaborative Learning Environments* (pp. 47–78). Springer Netherlands. https://doi.org/10.1007/978-94-017-0781-7_3.

Baker, M., Hansen, T., Joiner, R., & Traum, D. (1999). The role of grounding in collaborative learning tasks. *Collaborative Learning: Cognitive and Computational Approaches, 31*, 63.

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

Brennan, K., & Resnick, M. (2012). *New frameworks for studying and assessing the development of computational thinking*. 1–25.

Buder, J. (2011). Group awareness tools for learning: Current and future directions. *Computers in Human Behavior, 27*(3), 1114–1117.

Chanel, G., Kivikangas, J. M., & Ravaja, N. (2012). Physiological compliance for social gaming analysis: Cooperative versus competitive play. *Interacting with Computers, 24*(4), 306–316.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology, 76*(6), 893–910. https://doi.org/10.1037/0022-3514.76.6.893.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7.

Cooper, M. A. (1999). Classroom choices from a cognitive perspective on peer learning. In *Cognitive perspectives on peer learning* (pp. 215–233). Lawrence Erlbaum Associates Publishers.

Damon, W. (1984). Peer education: The untapped potential. *Journal of Applied Developmental Psychology, 5*(4), 331–343. https://doi.org/10.1016/0193-3973(84)90006-6.

Damon, W., & Phelps, E. (1989). Critical distinctions among three approaches to peer education. *International Journal of Educational Research, 13*(1), 9–19. https://doi.org/10.1016/0883-0355(89)90013-X.

Dawson, M. E., Schell, A. M., & Filion, D. L. (2007). The electrodermal system. *Handbook of Psychophysiology, 2*, 200–223.

Dede, C. (2010). Comparing frameworks for 21st century skills. *21st Century Skills: Rethinking How Students Learn, 20*, 51–76.

Elkins, A. N., Muth, E. R., Hoover, A. W., Walker, A. D., Carpenter, T. L., & Switzer, F. S. (2009). Physiological compliance and team performance. *Applied Ergonomics, 40*(6), 997–1003.

Garbarino, M., Lai, M., Bender, D., Picard, R. W., & Tognetti, S. (2014). *Empatica E3—A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition*. 39–42.

Haataja, E., Malmberg, J., & Järvelä, S. (2018). Monitoring in collaborative learning: Co-occurrence of observed behavior and physiological synchrony explored. *Computers in Human Behavior, 87*, 337–347. https://doi.org/10.1016/j.chb.2018.06.007.

Hadwin, A. F., Järvelä, S., & Miller, M. (2011) Self-regulated, co-regulated, and socially shared regulation of learning. In *Handbook of self-regulation of learning and performance* (pp. 65–84). Routledge/Taylor & Francis Group.

Henning, R. A., Boucsein, W., & Gil, M. C. (2001). Social–physiological compliance as a determinant of team performance. *International Journal of Psychophysiology, 40*(3), 221–232.

Henning, R., Armstead, A., & Ferris, J. (2009). Social psychophysiological compliance in a four-person research team. *Applied Ergonomics, 40*(6), 1004–1010.

Järvelä, S., Kivikangas, J. M., Kätsyri, J., & Ravaja, N. (2014). Physiological linkage of dyadic gaming experience. *Simulation & Gaming, 45*(1), 24–40.

Martinez Maldonado, R., Kay, J., Yacef, K., & Schwendimann, B. (2012). An interactive Teacher's dashboard for monitoring groups in a multi-tabletop learning environment. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent tutoring systems* (pp. 482–492). Berlin Heidelberg: Springer.

Meier, A., Spada, H., & Rummel, N. (2007). A rating scheme for assessing the quality of computer-supported collaboration processes. *International Journal of Computer-Supported Collaborative Learning, 2*(1), 63–86. https://doi.org/10.1007/s11412-006-9005-x.

Montague, E., Xu, J., & Chiou, E. (2014). Shared experiences of technology and trust: An experimental study of physiological compliance between active and passive users in technology-mediated collaborative encounters. *IEEE Transactions on Human-Machine Systems, 44*(5), 614–624.

Parkinson, B., & Simons, G. (2009). Affecting others: Social appraisal and emotion contagion in everyday decision making. *Personality and Social Psychology Bulletin, 35*(8), 1071–1084. https://doi.org/10.1177/0146167209336611.

Pijeira-Díaz, H. J., Drachsler, H., Järvelä, S., & Kirschner, P. A. (2016). *Investigating collaborative learning success with physiological coupling indices based on electrodermal activity*. 64–73.

Roschelle, J. (1992). Learning by collaborating: Convergent conceptual change. *The Journal of the Learning Sciences, 2*(3), 235–276.

Salomon, G., & Globerson, T. (1989). When teams do not function the way they ought to. *International Journal of Educational Research, 13*(1), 89–99. https://doi.org/16/0883-0355(89)90018-9 Q5

Schneider, B. (2019). Unpacking collaborative learning processes during hands-on activities using Mobile eye-tracking. In *Proceedings of 13th International Conference of Computer Supported Collaborative Learning* (pp. 156–165).

Schneider, B., & Pea, R. (2013). Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *International Journal of Computer-Supported Collaborative Learning, 8*(4), 375–397. https://doi.org/10.1007/s11412-013-9181-4.

Stahl, G. (2007). Meaning making in CSCL: Conditions and preconditions for cognitive processes by groups. *Proceedings of the 8th Iternational Conference on Computer Supported Collaborative Learning*, 652–661.

Starr, E., Dich, Y., Reilly, J., & Schneider, B. (2018). Using multi-modal learning analytics to support and measure collaboration in co-located dyads. In *12th international conference of the learning sciences*. London: UK.

Taylor, S., Jaques, N., Chen, W., Fedor, S., Sano, A., & Picard, R. (2015). *Automatic identification of artifacts in electrodermal activity data*. 1934–1937.

Wang, H., Chignell, M., & Ishizuka, M. (2006). Empathic tutoring software agents using real-time eye tracking. *Proceedings of the 2006 Symposium on Eye Tracking Research & Applications*, 73–78. https://doi.org/10.1145/1117309.1117346.

Webb, N. M., & Palincsar, A. S. (1996). Group processes in the classroom. Prentice Hall International.

Webb, N. M., Troper, J. D., & Fall, R. (1995). Constructive activity and learning in collaborative small groups. *Journal of Educational Psychology, 87*(3), 406–423. https://doi.org/10.1037/0022-0663.87.3.406.

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education, 46*(1), 71–95. https://doi.org/10.1016/j.compedu.2005.04.003.

Weintrop, D., & Wilensky, U. (2015). *Using commutative assessments to compare conceptual understanding in blocks-based and text-based programs*. 11th Annual ACM Conference on International Computing Education Research, ICER 2015.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Bertrand Schneider[1] • Yong Dich[1] • Iulian Radu[1]

✉ Bertrand Schneider
bertrand_schneider@gse.harvard.edu

✉ Yong Dich
ydich@college.harvard.edu

Iulian Radu
iulian_radu@gse.harvard.edu

[1] Graduate School of Education, Harvard University, Cambridge, MA, USA