Focused or Stuck Together: Multimodal Patterns Reveal Triads' Performance in Collaborative Problem Solving

Hana Vrzakova
University of Colorado Boulder
Boulder, USA
hana.vrzakova@colorado.edu

Mary Jean Amon
University of Central Florida
Orlando, USA
mamon@ist.ucf.edu

Angela Stewart
University of Colorado Boulder
Boulder, USA
angela.stewart@colorado.edu

Nicholas D. Duran
Arizona State University
Glendale, USA
nicholas.duran@asu.edu

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

ABSTRACT

Collaborative problem solving (CPS) in virtual environments is an increasingly important context of 21st century learning. However, our understanding of this complex and dynamic phenomenon is still limited. Here, we examine *unimodal primitives* (activity on the screen, speech, and body movements), and their *multimodal combinations* during remote CPS. We analyze two datasets where 116 triads collaboratively engaged in a challenging visual programming task using video conferencing software. We investigate how UI-interactions, behavioral primitives, and multimodal patterns were associated with teams' subjective and objective performance outcomes. We found that idling with limited speech (i.e., silence or backchannel feedback only) and without movement was negatively correlated with task performance and with participants' subjective perceptions of the collaboration. However, being silent and focused during solution execution was positively correlated with task performance. Results illustrate that in some cases, multimodal patterns improved the predictions and improved explanatory power over the unimodal primitives. We discuss how the findings can inform the design of real-time interventions for remote CPS.

CCS CONCEPTS

Human-centered computing ~ Collaborative Content Creation; Computer Supported Cooperative Work; Empirical studies in collaborative and social computing.

KEYWORDS

Multimodal Learning Analytics; Interpretability; CSCL; CSCW

ACM Reference format:

Hana Vrzakova, Angela Stewart, Mary Jean Amon, Nicholas Duran, Sidney K. D'Mello. 2020. Focused or Stuck Together: Multimodal Patterns Reveal Triads' Performance in Collaborative Problem Solving. *LAK '20,* March 23–27, 2020, Frankfurt, Germany. ©2020 Association for Computing Machinery. ACM ISBN 978-1-4503-7712-6/20/03...$15.00 https://doi.org/10.1145/3375462.3375467

1 Introduction

"*What shall we do about it?*" The question was followed by uncomfortable silence. Nobody moved. Clearly the team was stuck on the problem and nobody had a clue how to move forward. "*How about if you run the simulation again and check the code step-by-step? You will easily spot your mistake*" is what a teacher might suggest. However,, no teacher was there. The team was alone and the clock was ticking.

Although remote collaborative problem solving (CPS) is less dramatic than described above, it shares many aspects of this situation. First, a teacher or facilitator is rarely present and, when they are present, they cannot attend to all students. Consequently, students working in remote teams on a given task cannot raise questions as in a traditional classroom and, thus, have to rely on available materials and other teammates. In addition, this already challenging situation is accentuated by the affordances of virtual environment, which are subpar compared to collocated interaction. Therefore, the team's success strongly depends not only on students' problem-solving skills but also on students' ability to collaborate with (often) strangers.

In addition, large-scale remote CPS sessions are challenging to assess from the perspective of learning analytics. Although it might be simple to characterize students' performance by concise measures such as time-to-task-completion and task score, these measures cannot describe the rich *multimodal* dynamic processes and interactions between teammates [6]. Therefore, it remains challenging to automatically pinpoint which team's actions and behaviors underline good performance and which reveal their shortcomings.

Prior research on understanding collaborative learning have relied on expert coding of video data [18,28]. However, expert coding is time-consuming, especially with large sample sizes. Consequently, current research has explored the potential of bottom-up data-driven approaches to complement expert coding and has expanded to the field of multimodal learning analytics [2]. Data-driven understanding of collaborative learning has advanced from the analyses of unimodal primitives, such as keystrokes and clickstreams, to the analyses of data streams obtained from the multiple sensors [39]. The main motivation has been that the use of multiple sensors and resources allows for holistic understanding of collaborative processes [29,41,53].

However, multimodal modeling often suffers from the lack of model interpretability [19,32,55]. That is, it is challenging to identify which behaviors, signals, patterns, or model parameters contribute to model predictions since model predictions are a result of all factors. To address this, we aim to identify interpretable patterns in teams' verbal and nonverbal behaviors that correlate with meaningful outcomes during remote CPS.

1.1 Contribution and research questions

We study students' performance in remote CPS through the lens of teams' interaction in a virtual environment and their verbal and non-verbal behaviors. Concretely, we investigate how *unimodal primitives* (i.e. interaction in the virtual environment, students' speech, and the team's body movement) and *multimodal combinations* of these primitives are associated with both the team's objective task score and the self-assessed subjective perceptions of the collaboration. In addition, we answer the theoretical question of whether "more is better" and compare whether the multimodal patterns are more predictive of these outcomes compared to the unimodal primitives. We address the following specific research questions (RQs):

- RQ1: What behavioral patterns (in terms of the team's interaction, speech, and body movement) emerge during remote collaborative problem solving?
- RQ2: How do these patterns predict subjective and objective outcomes?
- RQ3: What is the advantage of multimodal patterns over unimodal primitives?

We address these questions in a novel research context. Whereas prior research has largely investigated dyads [36,42] or teams working in collocated settings [38,54], we use two large-scale datasets (348 students in total) [8,43] of triads who collaborated remotely over a shared virtual environment.

2 Background & Related Work

2.1 Process Gain and Loss during CPS

CPS occurs when two or more people engage in a coordinated attempt to solve a problem [37,59]. Intuitively, we may expect that multiple people working together on a task might achieve better outcomes, compared to an individual (this is called process gain [25]). However, that is often not the case. Groups often perform worse than they should because they engage in faulty CPS processes, a phenomenon known as process loss [16,25]. Process loss has been attributed to multiple factors, the most common being coordination losses, such as production blocking during collective ideation [31], the common-knowledge effect where there is an overemphasis on shared versus individual knowledge [13], and group-think where individual members converge to the dominant view [21]. Further, motivation losses, such as social-loafing [23], evaluation apprehension [5] and free-rider effect [24], further contribute to groups' underperformance.

Process loss gets amplified in remote collaborations, where collaborators do not have the rich of social signals available in face-to-face interactions [40]. Lagged, low quality, or non-existent audio and video channels dampen basic social signals [40]. Thus, process loss might be more severe in remote CPS.

2.2 Modeling Behavioral Patterns during Collaborations

Considerable work has been dedicated to data-driven modeling of collaboration. Data-driven approaches have mainly aimed to model low-level behaviors, such as turn-taking [3], joint attention [14,33], or synchrony and coordination [4,44]. Such behaviors have been modeled from speech [3,44], interaction patterns [3], eye gaze [33], physiology [34], and face or head pose [14,33].

Recent efforts have extended beyond low-level signals to model high-level collaborative behavioral patterns. For example, team management dialog has been modeled from eye gaze [22], social regulation from features of computer interaction [9], gender dynamics from language features [27], and collaboration quality from discourse features [17]. Specific to CPS, language-based features have been used to model facets of CPS (e.g. negotiating ideas) [15,45] and their corresponding behavioral indicators (e.g. asking for clarification) [10].

Unimodal features presumably cannot richly capture complex social interactions. Thus, multimodal signals have been increasingly used in modeling high-level collaborative patterns [1,44] as well as on students' states and traits, such as empathy [20], engagement [52], workload [26], and learning gains [35]. For example, Yoo and Kim [56] modeled project grades of long-term, online discussion groups using multimodal behavioral and linguistic patterns [56]. Follow-up analyses provided added interpretability to their results, providing the insight that acting as an information giver, using positive emotion words, and collaborating further in advance of the deadline (i.e. not procrastinating) positively related to project grades.

For CPS specifically, Murray and Oertel [30] modeled expert-rated task performance on a discussion-based CPS task. They trained a Random Forest classifier to predict task performance from acoustic-prosodic and linguistic features with a mean-squared error of 64.4 (baseline = 79.3). While a multimodal feature set did yield the best performance, it remained unclear which features precisely were the predictive. Other data-driven approaches, such as recurrence quantification analysis, have investigated collaborative learning from multiple modalities [1,7,8]. Despite these analyses work with complex systems of signals, the contribution of individual signals and modalities cannot be isolated from the results.

In summary, preliminary work has demonstrated the feasibility and utility of leveraging multimodal signals to predict team performance, more research is needed to understand the contribution of each modality to team processes and outcomes.

3 Data Collection

We used data from sources that were previously published elsewhere (dataset I [46]; dataset II [8]). The task in both datasets was similar with several nuances which we summarize below.

3.1 Participants

In dataset I, participants were 111 undergraduate students from a medium-sized private Midwestern university (63.1% female, average age = 19.4). Students were 74.8% Caucasian, 9.9% Hispanic/Latino, 8.1% Asian, 0.9% Black, 0.9% American Indian/Native Alaskan, 2.7% other, and 2.7% did not report ethnicity. Students were compensated with two hours of course credit. Prior to participation, students were asked to confirm that they had no previous experience with computer programming, which was the only inclusion criteria for this study. Students were assigned to 37 teams of three based on scheduling constraints. Thirty students from 10 teams indicated they knew at least one person in their team prior to participation.

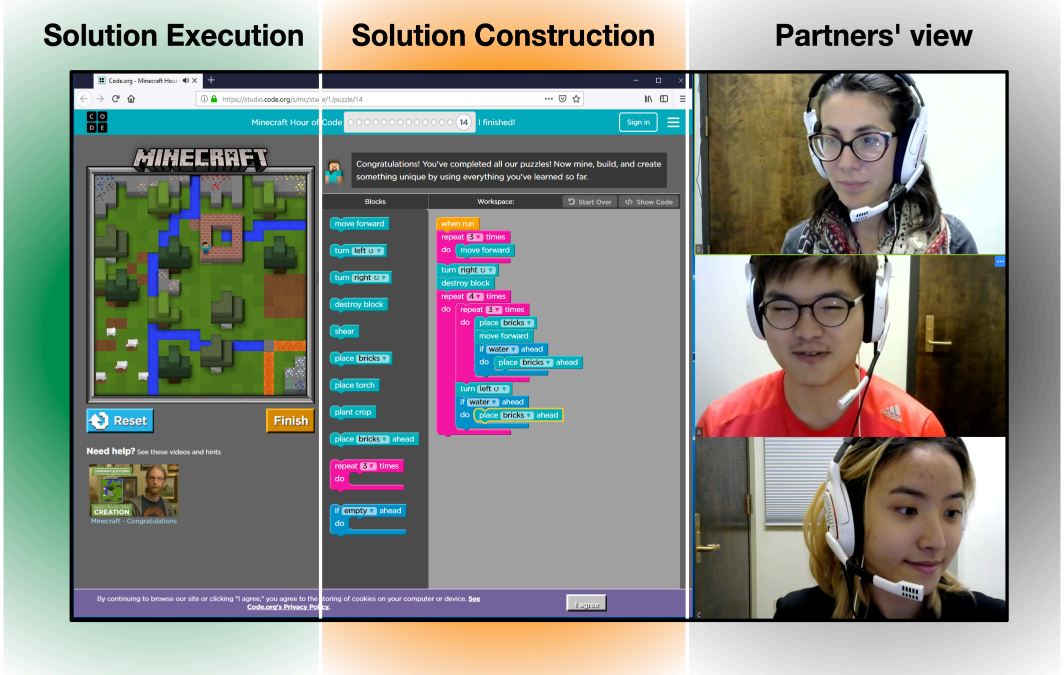
In dataset II, participants consisted of 303 students from two large public universities (56% female, average age = 22 years). Students were 47% Caucasian, 28% Hispanic/Latino, 18% Asian, 2% Black or African American, 1% American Indian or Alaska Native, and 4% other. Students were compensated either with a $50 gift card or with 3.5 hours of course credit. Prior to participation, students were asked to confirm that they met three inclusion criteria: 1) they spoke English, 2) they had no significant vision impairments, and 3) they had no prior experience with a physics game (unrelated to this study). Students were assigned to 101 teams of three based on scheduling constraints. Thirty students from 18 teams (26%) indicated they knew at least one person from their team prior to participation. Here, we use data from 116 teams – 32 teams from dataset I and 84 teams from dataset II.

3.2 Task Environment and CPS Task

We employed code.orgs's Minecraft-themed Hour of Code (Studio, 2014) as our CPS environment. Hour of Code is an online resource for students of all ages to learn basic computer programming principles in an hour. It employs a visual programming language, called Blockly [11], that represents lines of code (e.g. loops) as blocks that only interlock with other blocks in a syntactically correct manner. In Hour of Code, students use code to build structures and navigate around obstacles. At any point during code construction, students can run their solution and visualize the results in a preview window (see Figure 1).

3.3 Procedure

Students were randomly assigned to one of three separate, computer-enabled workstations in a lab. The workstations were either in separate rooms or partitioned in the same room with dividers. Each computer was equipped with a webcam and headset microphone for video conferencing with screen-sharing through Zoom (https://zoom.us). The headset microphone recorded student speech at either 16000 Hz (dataset I) or a variable frame rate (dataset II). An additional webcam was used to record the student's face and upper body at 10 Hz (dataset I) or a variable



Congratulations! You've completed all our puzzles! Now mine, build, and create something unique by using everything you've learned so far.

```
when run
repeat 3 times
do  move forward
turn right
destroy block
repeat 4 times
do  repeat 3 times
    do  place bricks
        move forward
        if water ahead
        do  place bricks ahead
    turn left
    if water ahead
    do  place bricks ahead
```

By continuing to browse our site or clicking "I agree," you agree to the toring of cookies on your computer or device. See Code.org's Privacy Pol cy.

Figure 1 User interfaces in Minecraft's Hour of Code. Students first constructed the code using the interlocking blocks of code (orange) and execute the code which ran the simulation in the Minecraft world (green). Students were equipped with microphones and headphones and could see each other in the partners' view (grey).

framerate (dataset II). Screen content was recorded using Zoom's built-in features at 25 Hz (dataset I) or at 5 Hz with custom screen recording software (dataset II).

Prior to engaging in the CPS task, students trained as a team in the lab (dataset I) or individually at home (dataset II) on how to use the Hour of Code environment. In this training, students completed five levels and viewed three accompanying videos that taught basic computer programming principles, such as loops and if statements.

The team was tasked with constructing a code that satisfied five criteria: 1) build a four-by-four brick building, 2) use at least one if-statement, 3) use at least one repeat loop, 4) build at least three bricks of the building over water, and 5) use 15-blocks of code or less. One randomly chosen student controlled interaction with the environment using the mouse and the other two students contributed to the solution. The task was time-constrained to 20 minutes for dataset I and 15 minutes for dataset II.

After completing the task, students *individually* rated their perception of the team. In dataset I, students were asked to rate their team's performance, communication, cooperation, and agreeableness using a six-point Likert scale (1 = very dissatisfied, 6 = very satisfied). In dataset II, students used a six-item questionnaire that assessed the quality of CPS processes. The questionnaire was based on a validated competency model of CPS [48] and assessed perception of the following CPS subfacets: sharing understanding of problems and solutions, establishing common ground, responding to others' questions and ideas thoughtfully, monitoring execution, fulfilling individual roles on the team, and taking initiative to advance the collaboration process. This was followed by a three-item inclusiveness and team norms questionnaire that assessed how inclusive the team was and whether the team worked towards task-related or socially-oriented goals [12]. Both the perceived CPS quality measures and the inclusiveness and team norms questionnaire were rated on a seven-point Likert scale (1 = disagree strongly, 7 = agree strongly). There were other CPS activities and assessments not germane to the present study and are not discussed further.

3.4 Outcome Measures

Each team's final solution was scored based on the five task criteria. Each criterion was worth one point, with final scores ranging from zero to five (M = 2.88, SD = 1.16).

In addition to objective performance (*task score*), we calculated a measure of subjective perception of the task (*subjective score*). For dataset I, we used individual self-reports of the team's performance, communication, cooperation, and agreeableness. We averaged measures of communication, cooperation, and agreeableness because ratings were highly correlated (*Cronbach's alpha* = .89). The averages were first computed per individual and, then, averaged across the three team members to obtain one score per team. Since perceptions of performance and collaboration were correlated (*Spearman's r* = 0.51), we averaged these two measures to yield a single subjective measure.

In dataset II, a subjective score was aggregated from the CPS quality and inclusiveness and team norms measures. We averaged six CPS items to yield a single score. This was also done for the inclusiveness and team norms measure. These two measures were highly correlated (Pearson's r = .79), so we combined them by first z-scoring each measure and, then, by averaging the z-scores. This was done first per individual and, then, averaged to the team level.

4 Identifying Unimodal Primitives and Multimodal Patterns

We analyze three modalities: 1) interaction in the virtual environment, 2) face and upper body movements, and 3) speech rate. We first preprocessed and unified sampling rates of all signals. Since the turns between students were quite short (*median* of 1.4 seconds) [8], we resampled the signals to 1Hz.

4.1 Behavioral Signals

We used the screen recording to extract a high-level measure of activity in the virtual environment since log files were unavailable. We focused on two areas of interest (AOI): solution construction and solution execution (Figure 1). We used a validated motion estimation algorithm [51] to compute the proportion of screen change in each AOI. Change in the solution construction AOI indicates solution edits, whereas change in the solution execution AOI indicates a team's attempt to test their code. We downsampled these time series to 1 Hz to ensure the same frequency across modalities. This was done by computing the mean of each AOI time series across non-overlapping 1-sec windows. The active AOI in each 1-sec window was identified as the one with the maximal proportion of pixels changed.

We computed a frame-level measure of face and upper body movement using the same validated motion estimation algorithm used for the screen AOIs. For data set I, the 10 Hz time series was transformed to a 1 Hz time series by taking the mean over non-overlapping 1-sec windows. For dataset II, face and upper body videos were recorded at a variable frame rate. We converted them to a constant frame rate of 10 frames per second using FFMPEG. We then converted the video to a 2 Hz time series by taking the mean over non-overlapping 0.5 second time windows. Finally, we converted the data to a 1 Hz time series by taking the mean over 1-second windows.

We computed speech rate as a measure of verbal participation in the task. We used the IBM Watson Speech to Text service [58] to generate transcriptions of individual audio recordings. IBM Watson provides start and stop times for each word spoken during the collaboration. For each second of the recording, we counted the number of words spoken during that second to yield speech rate (words per second). If a word spanned multiple seconds, we assigned it to the second in which it started.

4.2 Identifying Unimodal Primitives

For each modality, we defined and extracted three basic patterns of activity (unimodal primitives). First, we identified three primitives based on activity in the virtual environment. Specifically, *Solution construction* corresponded to the activity in the code region, whereas *Solution execution* represented activity in the Minecraft simulation region, in which students were running the assembled code. If changes occurred in both areas, we

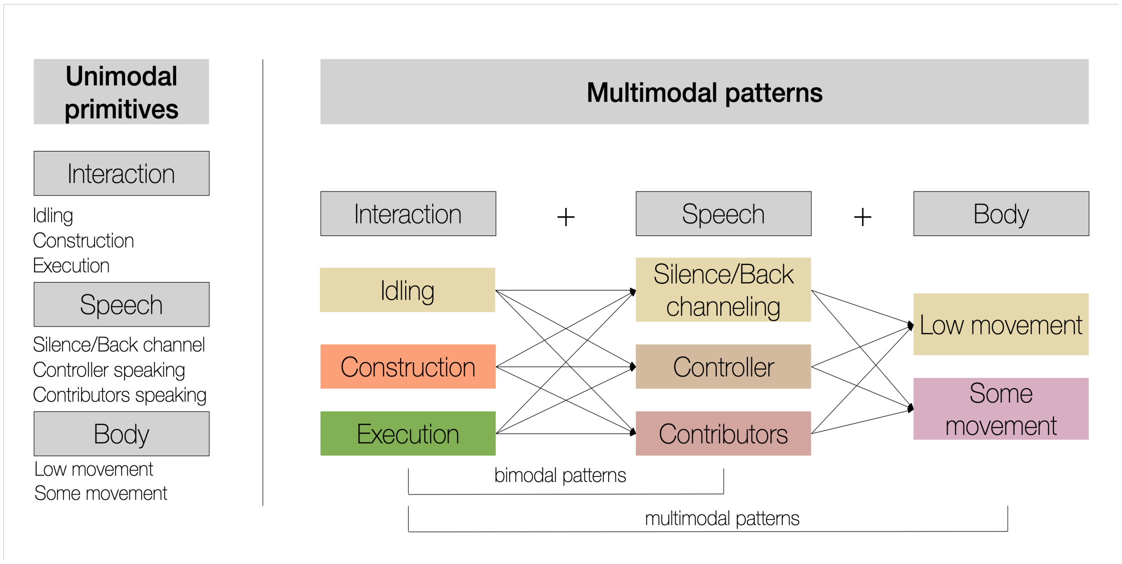


Figure 2 Overview of unimodal primitives and their bimodal and multimodal patterns.

selected the one with the most change as the dominant AOI for that second. A third primitive called *Idling* measured students' inactivity in the virtual environment. We identified idling based on a threshold ($t = 0.001\%$) [1], which accounted for small changes not associated with the students' actions (e.g., an icon randomly blinked in the Minecraft simulation).

Next, we computed three primitives for the team's speech activity. *Controller speaking* reflects instances when the student who was controlling the Minecraft environment was speaking. *Contributors speaking* reflects when either of the other two students were speaking. *Silence/Back channeling* represents the periods of interaction when the controller's and contributors' speech rate was below their median. In the case of the controller, the median was equal to 0. In the case of the contributors, the median often varied between 0 and 3 words, corresponding to either silence or back channeling (i.e., "uh-huh", "okay", "right"). To account for individual differences, the speech rate signals were first z-score standardized for each student. In addition, the two contributors' speech rates were averaged since it is not theoretically interesting to distinguish among the two. Then, we binarized the controller's and contributors' data streams based on their medians and identified which of three speech primitives occurred in each 1-second segment.

Finally, body-movement primitives were identified from the students' proportions of pixel changes in the video recording of facial expressions and upper body movements. The data stream of students' body movements was the noisiest of the three modalities considered. Even though a student might appear to be calm and focused, they could still exhibit small movements and gestures, such as scratching their chin with a pencil or fidgeting on the chair. Thus, we opted for two simple primitives at the team level. *Some movement* represented the moments when anybody in the team exhibited movement above their median, whereas *low movement* reflected the collective lack of body movement. As with speech rate, the body movement data streams were first z-score standardized for each student in the team separately, and averaged contributors' data streams after that. Next, two data streams (controller's and contributors') were binarized based on their median and then each second was classified as *Low movement* (all <= median) or *Some movement* (either controller's or contributors' values > median).

4.3 Combining Primitives into Multimodal Patterns

We hypothesized that combining modalities would provide added insights on the team's collaborative outcomes. Thus, we combined the unimodal primitives to generate multimodal patterns, as illustrated in Figure 2. First, we combined the interaction primitives with speech to yield nine bimodal patterns (3 for interaction × 3 speech). Then, we combined these with the body-movement primitives to yield 18 multimodal patterns (3 for interaction × 3 speech × 2 body movement). It should be noted that we consider bodily movements a secondary channel, compared to speech and interaction, because the task is dependent on speech and interaction. Therefore, we proceeded in the manner described to ascertain if there were any added benefits of including secondary signals to more basic ones.

4.4 Aggregation & Standardization by Dataset and School

We separately calculated the proportion of bimodal and multimodal patterns within each team by averaging across the 1-sec segments. In total, the pool of patterns comprised proportions of seven unimodal primitives, nine bimodal, and 18 multimodal patterns. We z-score standardized the proportional occurrences of each primitive/pattern separately within dataset I and II. In dataset II, the primitives/patterns were also standardized based on the school to account for differences between two schools. We similarly z-score standardized two outcome variables (task score and the subjective perceptions score).

5 RESULTS

We present the results with respect to our three research questions.

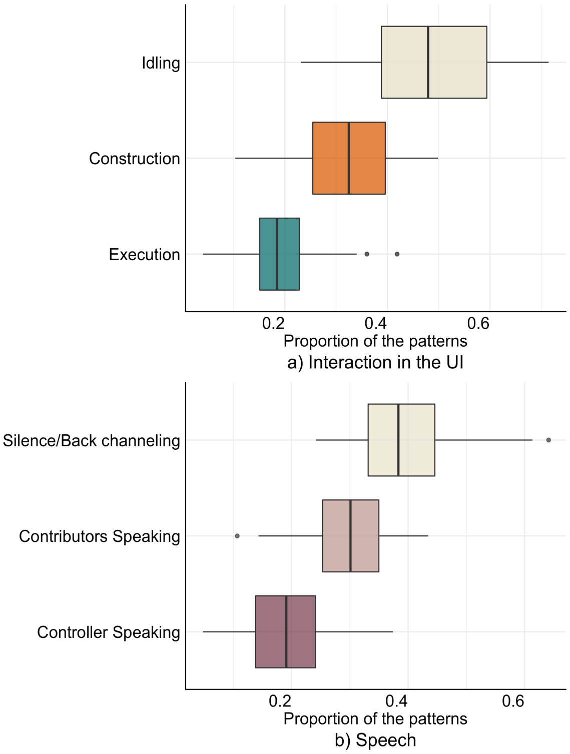
5.1 RQ1. Behavioral Patterns Emerging during Remote CPS

We found that *inactivity* in the user interface and speech represented the major unimodal primitives. On average, idling occurred 48.37% of the time (*SD* = 12.1) and silence/back channeling occurred 39.06% of the time (*SD* = 8.2). With respect to activity in the user interface, solution construction was the second most frequent (*M* = 32.67%, *SD* = 8.7%) followed by the solution execution (*M* = 18.96%, *SD* = 6.3). With respect to team's speech, contributors speaking (*M* = 29.77%, *SD* = 6.8) was more frequent than the controller speaking (*M* = 18.85%, *SD* = 7.4), which may indicate that the controller was often following the suggestions of the contributors. These data are shown in Figure 3. Since proportion of teams' body movements were equal because of the median split, they were omitted from the figure.

The bimodal and multimodal patterns provide a more detailed picture (Figure 4). For example, solution construction while silence/back channeling and without movement occurred more frequently (*M* = 10.45%, *SD* = 3.2) than the equivalent combination with the contributors (*M* = 6.03%, *SD* = 2.5) or the controller speaking (*M* = 3.96%, *SD* = 2.3). While it would be interesting to explore particular multimodal patterns in an of themselves, we mainly focus on the ones associated with the team's performance, which we analyzed next.

5.2 RQ2. Correlations with Team-level Outcomes

We first correlated the unimodal patterns with the objective task score and subjective score using Spearman correlation to address nonnormal distributions. We did not apply a correction for multiple significance tests due to the exploratory nature of this research and because false positives can be easily detected. For example, a significant correlation with a bimodal pattern where the corresponding unimodal or multimodal correlations were non-significant is likely a false positive. Tables 1 and 2 illustrate



a) Interaction in the UI

b) Speech

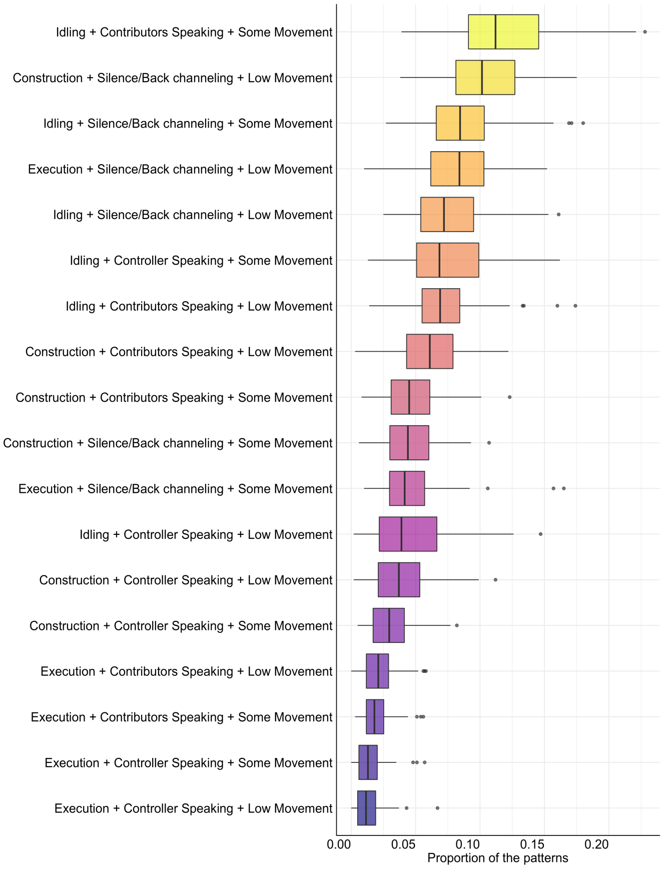


Figure 3 and 4 Distribution of unimodal primitives (left) and multimodal patterns (right). Proportions have been sorted in the descending order according to pattern's average.

correlations between performance and the interaction primitives (the speech primitives were uncorrelated with performance and are not shown) along with corresponding bimodal (speech+interaction) and multimodal patterns.

We found that task score was positively correlated with solution execution ($r = 0.334$, $p < 0.001$), suggesting that teams' efforts to try various solutions was positively related to their performance. When examining this primitive further, the bimodal and multimodal patterns revealed that solution execution during silence/back channeling ($r = 0.329$, $p < 0.001$) and with less movement ($r = 0.340$, $p < 0.001$) was similarly correlated with the task score. However, inclusion of some other primitives reduced the correlation (e.g., solution execution in silence/back channeling, but with some movement; $r = 0.203$, $p = 0.029$) and some others eliminated it altogether (e.g., solution execution when contributors speaking with some movement; $r = 0.106$, $p = 0.260$). The results suggest that even small changes in the context, such as contributors speaking or somebody moving, lowered the correlation of the pattern with the task score.

Results also suggest that inactivity was generally negatively associated with task score. Specifically, idling, as a unimodal primitive, was negatively correlated with the task score ($r = -0.204$, $p = 0.028$) as was idling in silence/back channeling ($r = -0.208$, $p = 0.025$). More importantly, the negative correlation was notably stronger when a lack of movement was added to the mix ($r = -0.351$, $p < 0.001$ for idling + silence/back channeling + low movement).

A similar picture unfolded with respect to the subjective score. The primitive pattern of solution execution was positively correlated with the subjective score ($r = 0.205$, $p = 0.027$). This correlation was strengthened when the contributors were speaking ($r = 0.273$, $p = 0.003$) and the team remained still ($r = 0.299$, $p = 0.001$). Similar to above, both idling ($r = -0.092$, $p = 0.326$) and idling in silence/back channeling ($r = -0.130$, $p = 0.165$) were

Table 1 and 2 Spearman's correlation of unimodal primitives, bimodal and multimodal patterns with team's task score (top) and subjective score (bottom). The asterisks indicate significance at the level of 0.05, 0.01 and 0.001 respectively.

Unimodal Primitives	Task score	Bimodal Patterns	Task score	Multimodal Patterns	Task score
Idling	-0.20*	Idling + Silence/Back channeling	-0.21*	Idling + Silence/Back channeling + Low Movement	-0.35***
				Idling + Silence/Back channeling + Some Movement	-0.06
		Idling + Controller Speaking	-0.09	Idling + Controller Speaking + Low Movement	-0.15
				Idling + Controller Speaking + Some Movement	-0.04
		Idling + Contributors Speaking	-0.11	Idling + Contributors Speaking + Low Movement	-0.11
				Idling + Contributors Speaking + Some Movement	-0.1
Construction	0.01	Construction + Silence/Back channeling	-0.07	Construction + Silence/Back channeling + Low Movement	-0.03
				Construction + Silence/Back channeling + Some Movement	-0.13
		Construction + Controller Speaking	0	Construction + Controller Speaking + Low Movement	0.01
				Construction + Controller Speaking + Some Movement	0.01
		Construction + Contributors Speaking	0.02	Construction + Contributors Speaking + Low Movement	0.07
				Construction + Contributors Speaking + Some Movement	-0.05
Execution	0.33***	Execution + Silence/Back channeling	0.33***	Execution + Silence/Back channeling + Low Movement	0.34***
				Execution + Silence/Back channeling + Some Movement	0.20*
		Execution + Controller Speaking	0.17	Execution + Controller Speaking + Low Movement	0.16
				Execution + Controller Speaking + Some Movement	0.15
		Execution + Contributors Speaking	0.18	Execution + Contributors Speaking + Low Movement	0.21*
				Execution + Contributors Speaking + Some Movement	0.11

Unimodal Primitives	Subjective score	Bimodal Patterns	Subjective score	Multimodal Patterns	Subjective score
Idling	-0.09	Idling + Silence/Back channeling	-0.13	Idling + Silence/Back channeling + Low Movement	-0.25**
				Idling + Silence/Back channeling + Some Movement	-0.04
		Idling + Controller Speaking	-0.01	Idling + Controller Speaking + Low Movement	-0.06
				Idling + Controller Speaking + Some Movement	0.02
		Idling + Contributors Speaking	0.01	Idling + Contributors Speaking + Low Movement	-0.03
				Idling + Contributors Speaking + Some Movement	0.05
Construction	-0.04	Construction + Silence/Back channeling	-0.05	Construction + Silence/Back channeling + Low Movement	0
				Construction + Silence/Back channeling + Some Movement	-0.15
		Construction + Controller Speaking	-0.06	Construction + Controller Speaking + Low Movement	-0.04
				Construction + Controller Speaking + Some Movement	-0.04
		Construction + Contributors Speaking	0.05	Construction + Contributors Speaking + Low Movement	0.13
				Construction + Contributors Speaking + Some Movement	-0.01
Execution	0.20*	Execution + Silence/Back channeling	0.12	Execution + Silence/Back channeling + Low Movement	0.12
				Execution + Silence/Back channeling + Some Movement	0.08
		Execution + Controller Speaking	0.07	Execution + Controller Speaking + Low Movement	0.02
				Execution + Controller Speaking + Some Movement	0.11
		Execution + Contributors Speaking	0.27**	Execution + Contributors Speaking + Low Movement	0.30**
				Execution + Contributors Speaking + Some Movement	0.18

negatively correlated with the subjective score (r = -0.254, p = 0.006).

There was also a notable lack of correlations. Specifically, none of the speech primitives correlated with team performance nor was solution construction and its associated bimodal and multimodal patterns. The action appears to lie in execution and idling.

5.3 RQ3. Contribution of Multimodal Patterns over Unimodal Primitives

We assessed whether the bimodal and multimodal patterns were more strongly correlated with the outcomes compared to the unimodal primitives. In case of the task score, the best correlation of 0.340 (p < 0.001) obtained with the Execution + Silence/Back channeling + Low movement multimodal pattern was similar to the correlation of 0.334 (p < 0.001) obtained via the unimodal primitive Execution. Zou's test of the difference between two overlapping dependent correlations [57] with one common variable (i.e. Execution) indicated that two correlation coefficients were statistically equivalent at p < 0.05 (CI [-0.10, 0.12]; i.e., the confidence interval overlaps with 0). Similarly for the subjective score, the best correlation of 0.205 (p = 0.027) obtained via the unimodal primitive Execution was statistically equivalent to the 0.299 correlation (p = 0.001) obtained from Execution + Contributors speaking + Low movement (CI [-0.08, 0.27]). The same trend was observed between unimodal and bimodal patterns. Thus, with respect to the solution execution, there was no added advantage of the bimodal or multimodal patterns over the unimodal primitives.

However, the patterns related to idling suggested a different conclusion. With respect to task score, the strongest negative correlation of -0.351 obtained with the multimodal pattern (Idling + Silence/Back channeling + Low movement) was statistically larger than the correlation of -0.204 (p = 0.028) from the unimodal primitive (Idling) ([-0.29, -0.00045]; CI does not overlap 0). This result was also found for the correlations with the subjective score. Specifically, the strongest negative correlation of -0.254 (p = 0.006), obtained with the same multimodal pattern, was statistically different ([-0.31, -0.01]) from the -0.092 (p = 0.326) correlation obtained with Idling alone. This finding suggests that not all idling is negatively associated with CPS. Idling while speaking and moving was not significantly related to the outcomes, but idling in silence or back channeling with little movement negatively predicted both objective and subjective outcomes.

6 Discussion

Multimodal learning analytics is gaining prominence in the field of collaborative learning. Researchers have typically explored data-driven approaches and multimodal modeling in an attempt to understand the rich and complex processes involved in collaboration. Despite the technical advances in machine learning, multimodal modeling often suffers from problems with model interpretability. In this work, we aimed to unveil complex, but interpretable, interaction and behavioral patterns that emerge during remote collaborative problem solving among triads.

6.1 Main Findings and Implications

We started with interaction patterns that emerged during CPS and gradually included primitives from speech and body movement with the goal of exploring how these patterns are associated with teams' subjective and objective performance. We found that certain patterns including code execution were positively correlated with teams' task and subjective score. Interestingly, the highest correlations were observed when the code execution occurred during periods of silence (or back channeling) and with little body movement, perhaps suggesting focused concentration. In contrast, idling with little speech and movement was negatively associated with both outcomes. This pattern might indicate that the team was stuck or experiencing a tense moment.

Surprisingly, the more infrequent unimodal primitive (code execution) was the one that was most strongly correlated with performance. Code execution was also correlated with task score when it was accompanied with silence or back channeling and little body movement. But even small changes in the context, such as teammates speaking or moving, weakened this association or eliminated it altogether. One possibility is that team's silence and stillness could indicate team's anticipation of a successful execution or their focused attention on the code itself as it was being executed. Additional "behavioral disturbances", such as teammates commenting on the code or fidgeting in their chairs, probably reduced teams' focus. Conversely, in the case of idling in the virtual environment, the negative correlation was stronger when idling was observed in the context of silence/back channeling and with less body movement and disappeared in the presence of speech.

These results lead us to question: are the multimodal patterns better than the unimodal primitives? As illustrated above, we found evidence for both sides of the argument. In the case of code execution, the answer is no, but it is a yes in the case of idling. However, it is important to go beyond the significant correlations as there is an informative signal in the non-significant ones as well. For example, consider idling once again. By itself, this pattern is negatively correlated with the task score (r = -.21) and the correlation is even more negative when idling is accompanied by silence/back channeling and little movement (r = -.35). However, there are many other configurations where idling is weak or negligible predictor of task score. For example, idling occurring in the context of the contributors speaking with some movement is more weakly correlated with task score (r = -.11) and the correlation is essentially null when idling is accompanied with the controller speaking and some movement (r = -.06). Thus, even when they do not improve predictive power, multimodal patterns help contextualize and reveal nuances in the unimodal primitives. This supports the overall idea of multimodal learning analytics in which the additional modalities (speech and body movement in our case) help to understand unclear patterns such as idling.

This finding is interesting from two perspectives. From the perspective of deeper understanding of collaborations , the multimodal patterns might be a preferable approach since they allow to identify contextual nuances in the collaborative process and, thus, increase interpretability. However, from the perspective of real-time interventions, the unimodal primitives represent a valuable source of information in and of themselves. In the design of learning interventions, one needs to compromise interpretability in favor of other factors, such as computational demands and intervention latency. As we further anticipate the advances in the field of multimodal modeling and pattern recognition, future research will need to investigate the trade-offs between the patterns' interpretability and usability for real-time use.

6.2 Limitations and Future Work

Like all studies, ours have limitations. First, we purposefully selected modalities that are already available at current PC setups (speech via a microphone, body movements via a webcam, interaction via a screen recording) and simplified their data streams into binary or ternary unimodal primitives. Although this approach greatly helps with interpretability, this comes at the cost of losing fine-grained detail. We also did not consider other modalities that might aid in interpretation (e.g., facial expressions).

Second, with respect to analyses, we opted for simpler approaches such as counting patterns that occurred simultaneously. However, additional approaches such as multi-dimensional recurrence quantification analyses [50] can be used to investigate the temporal dynamics of these patterns [1,8,49]. On that note, we also used zero-order Spearman correlations to study associations of the patterns with CPS outcomes. Although these analyses allowed us to highlight differences between patterns, they do not control for additional factors that could influence team performance. Future research should include factors such as the team's demographic composition, personality differences, prior knowledge effects, and task-related aspects (i.e., establishing common grounds, setting goals, or getting familiar with teammates) in order to study the incremental predictive validity of the patterns over these more stable factors.

Future work should also explore the patterns in relation to the theory of process loss. Although we did not directly test for specific effects, we hypothesized that process loss is an inevitable part of collaboration and would be reflected in some patterns. For example, a silent teammate could signal a lack of engagement and presumably signal a free rider effect. Similarly, increased speech could indicate a dominant teammate potentially blocking others. Further analyses could examine whether unimodal and multimodal patterns reflect these effects.

Finally, we explored the interaction and behavioral patterns in one task in a lab study. However, further research could examine generalizability of the constructs to other CPS tasks with data collected in more authentic environments.

6.3 Concluding Remarks

Remote collaborative problem solving is composed of rich multimodal dynamic processes and interactions between teammates that characterize team performance. Our work provides in-detailed insights on interaction patterns and behaviors observed in the large-scaled datasets and has implications for the design of real-time interventions.

Acknowledgments

We thank our colleagues, Lucca Eloy, Caroline Reinhardt, and Amanda Michaels, for their help with data collection and processing. The work was supported by National Science Foundation (NSF DUE 1745442) and Institute of Education Sciences (IES R305A170432). Any opinions, findings and conclusions or recommendations expressed in this material are those of authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] Mary Jean Amon, Hana Vrzakova, and Sidney K D'Mello. 2019. Beyond Dyadic Coordination: Multimodal Behavioral Irregularity in Triads Predicts Facets of Collaborative Problem Solving. *Cogn. Sci.* 43, 10 (2019).

[2] Paulo Blikstein. 2013. Multimodal learning analytics. In *ACM International Conference Proceeding Series.*

[3] Dan Calacci, Oren Lederman, David Shrier, and Alex "Sandy" Pentland. 2016. Breakout: An Open Measurement and Intervention Tool for Distributed Peer Learning Groups. *CoRR* abs/1607.0, (2016).

[4] Prerna Chikersal, Maria Tomprou, Young Ji Kim, Anita Williams Woolley, and Laura Dabbish. 2017. Deep Structures of Collaboration: Physiological Correlates of Collective Intelligence and Group Satisfaction. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17), 873–888.

[5] Michael Diehl and Wolfgang Stroebe. 1987. Productivity loss in brainstorming groups: Toward the solution of a riddle. *J. Pers. Soc. Psychol.* 53, 3 (1987), 497.

[6] Pierre Dillenbourg. 1999. What do you mean by collaborative learning?

[7] Muhterem Dindar, Iman Alikhani, Jonna Malmberg, Sanna Järvelä, and Tapio Seppänen. 2019. Examining shared monitoring in collaborative learning: A case of a recurrence quantification analysis approach. *Comput. Human Behav.* April 2018 (2019).

[8] Lucca Eloy, Angela E.B. Stewart, Mary Jean Amon, Caroline Reinhardt, Amanda Michaels, Chen Sun, Valerie Shute, Nicholas D Duran, and Sidney D&#39;Mello. 2019. Modeling Team-level Multimodal Dynamics During Multiparty Collaboration. In *2019 International Conference on Multimodal Interaction* (ICMI '19), 244–258. Retrieved from http://doi.acm.org/10.1145/3340555.3353748

[9] Abigail C Evans, Jacob O Wobbrock, and Katie Davis. 2016. Modeling Collaboration Patterns on an Interactive Tabletop in a Classroom Setting. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (CSCW '16), 860–871.

[10] Michael Flor, Su-Youn Yoon, Jiangang Hao, Lei Liu, and Alina von Davier. 2016. Automated classification of collaborative problem solving interactions in simulated science tasks. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 31–41.

[11] Neil Fraser. 2015. Ten things we've learned from Blockly. In *Proceedings of the 2015 IEEE Blocks and Beyond Workshop*, 49–50.

[12] Donald G Gardner and Jon L Pierce. 2016. Organization-based self-esteem in work teams. *Gr. Process. Intergr. Relations* 19, 3 (2016), 394–408.

[13] Daniel Gigone and Reid Hastie. 1993. The common knowledge effect: Information sharing and group judgment. *J. Pers. Soc. Psychol.* 65, 5 (1993), 959.

[14] Carl Gutwin, Scott Bateman, Gaurav Arora, and Ashley Coveney. 2017. Looking Away and Catching Up: Dealing with Brief Attentional Disconnection in Synchronous Groupware. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17), 2221–2235.

[15] Jiangang Hao, Lei Chen, Michael Flor, Lei Liu, and Alina A von Davier. 2017. CPS-Rater: Automated Sequential Annotation for Conversations in Collaborative Problem-Solving Activities. *ETS Res. Rep. Ser.* 2017, 1 (2017), 1–9.

[16] Tobias Hecking, Dorian Doberstein, and H Ulrich Hoppe. 2019. Predicting the Well-functioning of Learning Groups Under Privacy Restrictions. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (LAK19), 245–249. Retrieved from http://doi.acm.org/10.1145/3303772.3303826

[17] Tiffany Herder, Zachari Swiecki, Simon Skov Fougt, Andreas Lindenskov Tamborg, Benjamin Brink Allsopp, David Williamson Shaffer, and Morten Misfeldt. 2018. Supporting Teachers' Intervention in Students' Virtual Collaboration Using a Network Based Model. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (LAK '18), 21–25. Retrieved from http://doi.acm.org/10.1145/3170358.3170394

[18] Cindy E. Hmelo-Silver and Howard S. Barrows. 2008. Facilitating collaborative knowledge building. *Cogn. Instr.* (2008).

[19] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. *Conf. Hum. Factors Comput. Syst. - Proc.* (2019), 1–13.

[20] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Ryuichiro Higashinaka, and Junji Tomita. 2018. Analyzing Gaze Behavior and Dialogue Act During Turn-taking for Estimating Empathy Skill Level. In *Proceedings of the 2018 on International Conference on Multimodal Interaction* (ICMI '18), 31–39. DOI:https://doi.org/10.1145/3242969.3242978

[21] Irving Lester Janis. 1982. *Groupthink: Psychological studies of policy decisions and fiascoes.* Houghton Mifflin Boston.

[22] Patrick Jermann and Kshitij Sharma. 2018. Gaze as a Proxy for Cognition and Communication. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, 152–154.

[23] Steven J Karau and Kipling D Williams. 1993. Social loafing: A meta-analytic review and theoretical integration. *J. Pers. Soc. Psychol.* 65, 4 (1993), 681.

[24] Norbert L Kerr and Steven E Bruun. 1983. Dispensability of member effort and group motivation losses: Free-rider effects. *J. Pers. Soc. Psychol.* 44, 1 (1983), 78.

[25] Norbert L Kerr and R Scott Tindale. 2004. Group performance and decision making. *Annu. Rev. Psychol.* 55, (2004), 623–655.

[26] Charlotte Larmuseau, Pieter Vanneste, Piet Desmet, and Fien Depaepe. 2019. Multichannel Data for Understanding Cognitive Affordances During Complex Problem Solving. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (LAK19), 61–70. DOI:https://doi.org/10.1145/3303772.3303778

[27] Yiwen Lin, Nia Dowell, Andrew Godfrey, Heeryung Choi, and Christopher Brooks. 2019. Modeling Gender Dynamics in Intra and Interpersonal Interactions During Online Collaborative Learning. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (LAK19), 431–435. DOI:https://doi.org/10.1145/3303772.3303837

[28] Emma M. Mercier, Steven E. Higgins, and Laura da Costa. 2014. Different leaders: Emergent organizational and intellectual leadership in children's collaborative learning groups. *Int. J. Comput. Collab. Learn.* (2014). DOI:https://doi.org/10.1007/s11412-014-9201-z

[29] Daniele Di Mitri, Jan Schneider, Roland Klemke, Marcus Specht, and Hendrik Drachsler. 2019. Read between the lines: An annotation tool for multimodal data for learning. In *ACM International Conference Proceeding Series.*

[30] Gabriel Murray and Catharine Oertel. 2018. Predicting Group Performance in Task-Based Interaction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (ICMI '18), 14–20. DOI:https://doi.org/10.1145/3242969.3243027

[31] Bernard A Nijstad, Wolfgang Stroebe, and Hein F M Lodewijkx. 2003. Production blocking and idea generation: Does blocking interfere with cognitive processes? *J. Exp. Soc. Psychol.* 39, 6 (2003), 531–548.

[32] Xavier Ochoa, Nadir Weibel, Marcelo Worsley, and Sharon Oviatt. 2016. Multimodal learning analytics data challenges. In *ACM International Conference Proceeding Series.* DOI:https://doi.org/10.1145/2883851.2883913

[33] Kazuhiro Otsuka, Keisuke Kasuga, and Martina Köhler. 2018. Estimating Visual Focus of Attention in Multiparty Meetings Using Deep Convolutional Neural Networks. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (ICMI '18), 191–199. DOI:https://doi.org/10.1145/3242969.3242973

[34] Richard V Palumbo, Marisa E Marraccini, Lisa L Weyandt, Oliver Wilder-Smith, Heather A McGee, Siwei Liu, and Matthew S Goodwin. 2017. Interpersonal autonomic physiology: A systematic review of the literature. *Personal. Soc. Psychol. Rev.* 21, 2 (2017), 99–141.

[35] Joseph M Reilly and Chris Dede. 2019. Differences in Student Trajectories via Filtered Time Series Analysis in an Immersive Virtual World. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (LAK19), 130–134. DOI:https://doi.org/10.1145/3303772.3303832

[36] Joseph M Reilly, Milan Ravenell, and Bertrand Schneider. 2018. Exploring Collaboration Using Motion Sensors and Multi-Modal Learning Analytics. *Int. Educ. Data Min. Soc.* (2018).

[37] Jeremy Roschelle and Stephanie D Teasley. 1995. The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning*, 69–97.

[38] Bertrand Schneider and Paulo Blikstein. 2018. Tangible User Interfaces and Contrasting Cases as a Preparation for Future Learning. *J. Sci. Educ. Technol.* (2018). DOI:https://doi.org/10.1007/s10956-018-9730-8

[39] Jan Schneider, Dirk Börner, Peter Van Rosmalen, and Marcus Specht. 2015. Augmenting the senses: A review on sensor-based learning support. *Sensors (Switzerland)* (2015). DOI:https://doi.org/10.3390/s150204097

[40] Julian Schulze and Stefan Krumm. 2017. The "virtual team player": A review and initial model of knowledge, skills, abilities, and other characteristics for virtual collaboration. *Organ. Psychol. Rev.* 7, 1 (2017), 66–95. DOI:https://doi.org/10.1177/2041386616675522

[41] Daniel Spikol, Marcelo Worsley, Luis P. Prieto, Xavier Ochoa, M. J. Rodríguez-Triana, and Mutlu Cukurova. 2017. Current and future multimodal learning analytics data challenges. In *ACM International Conference Proceeding Series.* DOI:https://doi.org/10.1145/3027385.3029437

[42] Emma L Starr, Joseph M Reilly, and Bertrand Schneider. 2018. Toward Using Multi-Modal Learning Analytics to Support and Measure Collaboration in Co-Located Dyads. . International Society of the Learning Sciences, Inc.[ISLS].

[43] Angela Stewart and Sidney K D'Mello. 2018. Connecting the Dots Towards Collaborative AIED: Linking Group Makeup to Process to Learning. In *International Conference on Artificial Intelligence in Education*, 545–556.

[44] Angela E. B. Stewart, Zachary A. Keirn, and Sidney K D'Mello. 2018. Multimodal Modeling of Coordination and Coregulation Patterns in Speech Rate During Triadic Collaborative Problem Solving. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (ICMI '18), 21–30. DOI:https://doi.org/10.1145/3242969.3242989

[45] Angela E.B. Stewart, Hana Vrzakova, Chen Sun, Jade; Yonehiro, Cathlyn Adele Stone, D Duran, Nicholas, Valerie Shute, and Sidney K. D'Mello. 2019. I Say, You Say, We Say: Using Spoken Language to Model Socio-Cognitive Processes during Computer-Supported Collaborative Problem Solving. *Proc. ACM Hum. Comput. Interact.* 3, November (2019), 19.

[46] Angela Stewart and Sidney K D Mello. Connecting the Dots towards Collaborative AIED : Linking Group Makeup to Process to Learning.

[47] Code Studio. Code Studio.

[48] Chen Sun, Valerie J Shute, Angela Stewart, Jade Yonehiro, Nicholas Duran, and Sidney D'Mello. 2020. Towards a generalized competency model of collaborative problem solving. *Comput. Educ.* 143, (2020), 103672.

[49] Hana Vrzakova, Mary Jean Amon, Angela E B Stewart, and Sidney K D'Mello. 2019. Dynamics of Visual Attention in Multiparty Collaborative Problem Solving Using Multidimensional Recurrence Quantification Analysis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 342:1--342:14.

[50] Sebastian Wallot, Andreas Roepstorff, and Dan Mønster. 2016. Multidimensional Recurrence Quantification Analysis (MdRQA) for the analysis of multidimensional time-series: A software implementation in MATLAB and its application to group-level data in joint action. *Front. Psychol.* 7, (2016), 1835.

[51] Jacqueline Kory Westlund, Sidney K D'Mello, and Andrew M Olney. 2015. Motion Tracker: camera-based monitoring of bodily movements using motion silhouettes. *PLoS One* 10, 6 (2015), e0130293.

[52] M Worsley. 2018. (Dis)Engagement matters: Identifying efficacious learning practices with multimodal learning analytics. (2018), 365–369.

[53] Marcelo Worsley. 2014. Multimodal learning analytics as a tool for bridging learning theory and complex learning behaviors. In *MLA 2014 - Proceedings of the 2014 ACM Multimodal Learning Analytics Workshop and Grand Challenge, Co-located with ICMI 2014.*

[54] Bin Xie, Joseph M Reilly, Yong Li Dich, and Bertrand Schneider. 2018. Augmenting Qualitative Analyses of Collaborative Learning Groups Through Multi-Modal Sensing. . International Society of the Learning Sciences, Inc.[ISLS].

[55] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating how experienced UX designers effectively work with machine learning. *DIS 2018 - Proc. 2018 Des. Interact. Syst. Conf.* (2018), 585–596.

[56] Jaebong Yoo and Jihie Kim. 2014. Can online discussion participation predict group project performance? Investigating the roles of linguistic features and participation patterns. *Int. J. Artif. Intell. Educ.* 24, 1 (2014), 8–32.

[57] Guang Yong Zou. 2007. Toward using confidence intervals to compare correlations. *Psychol. Methods* 12, 4 (2007), 399.

[58] IBM. Retrieved May 2, 2018 from https://www.ibm.com/watson/services/speech-to-text/

[59] 2015. *PISA 2015 Collaborative Problem Solving Framework.*