Intern. J. Comput.-Support. Collab. Learn (2021) 16:485–523
https://doi.org/10.1007/s11412-021-09358-2

Many are the ways to learn identifying multi-modal behavioral profiles of collaborative learning in constructivist activities

Jauwairia Nasir[1] · **Aditi Kothiyal**[1] · **Barbara Bruno**[1] · **Pierre Dillenbourg**[1]

Received: 29 April 2021 / Accepted: 18 October 2021 / Published online: 21 January 2022

Abstract
Understanding the way learners engage with learning technologies, and its relation with their learning, is crucial for motivating design of effective learning interventions. Assessing the learners' state of engagement, however, is non-trivial. Research suggests that performance is not always a good indicator of learning, especially with open-ended constructivist activities. In this paper, we describe a combined multi-modal learning analytics and interaction analysis method that uses video, audio and log data to identify multi-modal collaborative learning behavioral profiles of 32 dyads as they work on an open-ended task around interactive tabletops with a robot mediator. These profiles, which we name *Expressive Explorers*, *Calm Tinkerers*, and *Silent Wanderers*, confirm previous collaborative learning findings. In particular, the amount of speech interaction and the overlap of speech between a pair of learners are behavior patterns that strongly distinguish between learning and non-learning pairs. Delving deeper, findings suggest that overlapping speech between learners can indicate engagement that is conducive to learning. When we more broadly consider learner affect and actions during the task, we are better able to characterize the range of behavioral profiles exhibited among those who learn. Specifically, we discover two behavioral dimensions along which those who learn vary, namely, problem solving strategy (actions) and emotional expressivity (affect). This finding suggests a relation between problem solving strategy and emotional behavior; one strategy leads to more frustration compared to another. These findings have implications for the design of real-time learning interventions that support productive collaborative learning in open-ended tasks.

Keywords Learning analytics · Collaborative learning · Engagement · Human-robot interaction · Multi-modal · Social robots

✉ Jauwairia Nasir
jauwairia.nasir@epfl.ch

Extended author information available on the last page of the article

Introduction

In line with constructivist learning theories, which suggest that it is important for learners to be engaged in constructing their own knowledge, there has been an increasing emphasis on incorporating more open-ended learning activities in classrooms (Hmelo-Silver et al., 2007). Such activities are learner-centered and revolve around an authentic problem or inquiry task which learners work on, often in groups. They are supported in these tasks by teachers and scaffolds within the learning environments (Hmelo-Silver et al., 2007; Land et al., 2000). The design of effective interventions to support learners as they work on such activities is crucial in order to have a positive impact on learning outcomes (Kirschner et al., 2006).

The increase in the use of technological tools such as tangible interfaces and robots in classrooms and recent advances in learning analytics have enabled building models of learners based on their actions in such technology-supported open-ended learning activities (Basu et al., 2017). These models can then be used to develop adaptive and/or personalized interventions to support learners in these open-ended activities. Learner models are usually built on some recognized notion of success or failure associated with each learner action (Desmarais & Baker, 2012). However, we argue that in open-ended learning activities where the goal is learner exploration, success can be a misleading measure of learning as the design can sometimes be based on the *productive failure* paradigm (Kapur, 2008). This paradigm proposes that getting learners to work on a complex problem-solving task before instruction where they are likely to fail can be conducive to learning. In fact, studies have shown that learners' success or failure on such activities does not impact learning outcomes (Loibl & Rummel, 2014). Therefore, systems that rely only on performance pose a risk of misjudging learners that are apparently failing in the task but end up with higher learning outcomes, as well as learners that do well in the task but do not exhibit learning (Do-lenh, 2012). Indeed, we also found that among learners who ended up with higher learning gains, there were some teams who actually failed in the task while, on the other hand, some teams that succeeded in the task did not learn (Nasir, Norman, Bruno, & Dillenbourg, 2020c). Similarly, in Do-lenh (2012), the authors compared the task performance and learning of logistics apprentices that interact with a tangible tabletop environment for warehouse manipulation versus those who use the traditional paper-and-pen based methods. They found that while the task performance was higher for the former, there was no increase in the learning gains relative to the traditional methods, which was attributed to over-engagement with the technology, without higher levels of reflection. The complex relationship between task performance and learning within open-ended learning activities thus makes it non-trivial to evaluate whether and when a learner is learning through inspection of their success in activities they engage in throughout their process. It is a challenge then to build comprehensive models of learners, including learners who truly learn despite their failures along the way during problem solving.

For effective collaborative learning to occur in open-ended learning environments, learners need to share and regulate their own and each other's cognition, metacognition, affect and motivation (Järvelä et al., 2020). This learning process is complex and its success has been evaluated based on indicators of discourse, gestures, gaze, cognition and social skills (Spikol et al., 2017; Stahl et al., 2013). Recent research has suggested that multi-modal data, i.e., integrating multiple of the behavioral indicators listed above, provides an opportunity to more comprehensively characterize learning in open-ended learning environments such as those involving engineering design (Blikstein & Worsley, 2016;

Spikol et al., 2017). We consider a behavior as *an action or expression (verbal or facial) of the learner while interacting with the learning environment or a team member*. Further, we refer to multi-modality as *the application and interplay of multiple semiotic modes in order to help understand a specific process, in this case, learning*. In previous work (Nasir, Bruno, et al., 2021a), we found that in an open-ended collaborative learning activity, multi-modal behaviors (extracted from video, audio, and log data) better distinguish those who learn from those who do not compared to when only a single modality was used. Further, we argue that it is not straightforward to classify a certain behavior as absolutely good or bad for learning. For example, D'Mello and Graesser (2012) propose a model to explain the dynamics of affective states that emerge during deep learning, that are also linked with cognitive engagement. Based on their studies, they suggest that frustration regulation in learners is important as frustration is considered a negative state (D'Mello & Graesser, 2012; Hone, 2006; Klein et al., 2002). On the other hand, Baker et al. (2010) suggest that boredom is more important than frustration. This is also supported by the work of Mentis et al. (2007) who propose that frustration only needs to be remediated when it occurs due to events that are not under the control of the user, for example, a system bug. Similarly, the literature on learning through failure suggests that "productive confusion" is conducive to learning as it enables learners to become aware of knowledge gaps and identify deep problem features (Lodge et al., 2018; Loibl et al., 2017).

The findings above together suggest that there is an interplay between behaviors and their associated states in their effect on collaborative learning; specifically, the role of a an affective state in collaborative learning depends on the context and the accompanying behaviors and states. This points to the need to examine multiple behaviors together to build a more robust understanding of learning, instead of attributing significance to individual behaviors. This is especially important when the goal is to detect where to intervene and scaffold learners appropriately during an activity. This motivates us to explore the use of multi-modal behavioral data to build comprehensive learning vs non-learning profiles in an open-ended collaborative learning setting. In this paper, we present an approach for identifying the *collection* of behaviors associated with learning. Specifically, we consider the corpus of multi-modal behavioral data collected during *JUSThink* (Nasir, Norman, Bruno, & Dillenbourg, 2020c), which by design follows the *problem-based learning* paradigm (Barron et al., 1998). Our goal in this work is to explore the role of multi-modality and identify specifically the collection of multi-modal behaviors that distinguish learning and non-learning groups. We argue that a collection of multi-modal behaviors may offer a richer characterization of collaborative learning in an open-ended activity, so that we may then use the identified learning profiles to build real-time robot interventions that are capable of scaffolding learners.

Formally, we investigate the following research question:

RQ: What do learners' visible behavior profiles reveal about learning in a collaborative open-ended learning activity?

Literature review

Research on problem-based learning suggests that learners collaboratively working on authentic, open-ended problems is effective for conceptual understanding (Barron et al., 1998; Kirschner et al., 2011). Furthermore, impasses have been shown to play an important

role in learning (Kapur, 2008; Schwartz & Bransford, 1998; Schwartz & Martin, 2004; VanLehn et al., 2003); for instance, during coached problem solving, learning seldom happens unless learners reach an impasse (VanLehn et al., 2003). Similarly, when learners solve authentic, open-ended problems collaboratively they often fail, but this failure is productive for learning and leads to deep conceptual understand ing and improved transfer (Kapur, 2008; Schwartz & Martin, 2004). Therefore, in this work, we broadly adopt the impasse-driven theories of learning such as productive failure, which suggest that

- Performance in problem solving is not necessarily an indicator of learning (Loibl & Rummel, 2014).
- Learning is driven by the mechanisms of becoming aware of ones' knowledge gaps, followed by recognition of deep knowledge structures engendered in moments of failure during problem solving (Lodge et al., 2018; Loibl et al., 2017).
- Learning while working on activities collaboratively and encountering failures requires learners to sustain and regulate their own and the teams' cognition, meta-cognition, emotions and behaviors towards completing the task and learning through impasses (Järvelä et al., 2020).

The theory above highlights the need to identify the multiple constructs that are together responsible for the success of collaborative impasse-driven learning. The effectiveness of collaborative learning depends on many factors, such as team members' speech, their actions within the learning environment, and their eye gaze (Spikol et al., 2017; Stahl et al., 2013). Further, in impasse-driven learning paradigms, as learners work on complex problems, there is a "zone of optimal confusion" (Lodge et al., 2018) where learners become aware of their knowledge gaps and subsequently recognize the deep features of the underlying concept (Loibl et al., 2017). In this zone, confusion can be productive. However, if learners' confusion persists, it can become unproductive and lead to frustration and then disengagement (D'Mello & Graesser, 2012). Thus, the regulation of emotions becomes crucial in impasse-driven learning situations to ensure that learners do not transcend into disengagement. Putting these factors together we argue that learning while collaborating in a technology-based open-ended activity depends on sharing and regulating learners *speech*, *actions*, *gaze* and *emotions*. Based on this theoretical framing, we choose to focus on these four indicators and their interplay to characterize collaborative learning. Below we elaborate on the literature related to the effect of each of these indicators on collaborative learning and then argue why it is necessary to integrate these indicators to build comprehensive profiles.

Indicators of collaborative learning

While collaboration can make learning more effective, especially in open-ended learning activities, several researchers stress that this depends on the quality of the interaction. Dillenbourg et al. (2009) emphasize that in collaborative settings, particular forms of interactions among people, such as productive verbal elaborations, are expected to occur, which could trigger learning mechanisms, but there is no guarantee that the expected interactions will actually occur. Other work (Barron, 2003; Lou et al., 2001; Meier et al., 2007) similarly suggests that the conditions under which collaborative learning is effective are diverse and complex. Hence, researchers have attempted to understand the collaborative learning mechanisms using various indicators of collaboration, such as learners' speech

(for instance, (Weinberger & Fischer, 2006)), eye gaze (Jermann & Nüssli, 2012), physiological measures (Schneider et al., 2020) and actions (Popov et al., 2017). From this data, they have identified conditions for productive collaborative learning. Below we describe some of these indicators and their relationship to productive collaborative learning.

Speech Speech plays a very important role in collaborative learning as it is primarily through dialogue that learners build a joint understanding of the shared problem space and engage in knowledge construction (Barron, 2003; Roschelle & Teasley, 1995; Teasley, 1997). Within learner dialogue (speech or chats), it has been found that the quantity (e.g., number and length of utterances, and talk time) and heterogeneity and transactivity of verbal participation (e.g., turn taking and building on each other's reasoning), along with features of speech, such as voice inflection, are all indicative of good collaboration (Martinez et al., 2011; Reilly & Schneider, 2019; Viswanathan & VanLehn, 2017; Weinberger & Fischer, 2006). Pauses are also considered an essential part of speech and dialogue as sometimes one pauses to breathe, to plan, or to check whether someone else wants to speak (Fors, 2015; Maroni et al., 2008). Research has shown that shorter pauses (200-500 ms), relative to longer pauses (>1000 ms), tend to be linked with positive perception of speech, the ease of understanding speech, as well as memorability (Fors, 2015). All of these aspects help with better communication, that is related to better collaboration and learning.

Eye gaze Eye gaze has been used, often along with dialogue, to evaluate collaborative learning (Jermann et al., 2011; Schneider & Pea, 2013; Sharma et al., 2021). Research has shown that measures of joint visual attention, such as cross-recurrence (Jermann et al., 2011; Jermann & Nüssli, 2012; Schneider et al., 2016) and gaze similarity (Sharma et al., 2015; Sharma et al., 2021) are related to increased collaboration quality and learning outcomes. On the other hand, a measure of gaze dispersion is found to be related to misunderstandings (Cherubini et al., 2008) and unbalanced gaze participation is negatively correlated with learning outcomes (Schneider et al., 2018). Similarly, sharing gaze among collaborators is related to improved collaboration (e.g., improved transactivity in learner dialogue) and learning gains (Schneider & Pea, 2013, 2015).

Actions Interaction logs within technology-enhanced learning environments are used to examine the state of learners' performance and learning in both individual and collaborative conditions. In collaborative learning, learners' clickstream or touch traces are used, often along with their dialogue, to identify productive actions and patterns (Evans et al., 2016; Martinez-Maldonado et al., 2013; Popov et al., 2017; Rodríguez & Boyer, 2015; Viswanathan & VanLehn, 2017). Research has shown that analytics of task-specific actions when learners collaborate in complex problem-solving environments can be used to distinguish high and low performers in collaborative learning (Emara et al., 2018; Kapur, 2011; Perera et al., 2008). For instance, while collaborating around an interactive tabletop, while the number or symmetry of actions and speech of each member of a team were not found to relate to collaboration quality, certain sequences of actions and speech were found to be indicative of quality of collaboration (Martinez-Maldonado et al., 2013). Specifically, low collaborating groups were found to act in parallel, without discussing, while high collaborating groups were found to work together on task-related objects while discussing. Other work has found that the combination of touches to unrelated objects on the screen and multiple users interacting with the screen at the same time can predict collaboration quality (Evans et al., 2016). However, in a chat-based collaborative learning environment,

researchers (Popov et al., 2017) found that neither alignment of learner actions (synchrony) nor learners building on each others' reasoning (transactivity) was related to performance on the task. On the other hand, other factors such as group dynamics and prior knowledge played a more critical role. Thus, the role of symmetry, synchrony and transactivity in actions during collaborative learning appears to depend on the context.

Affect Affect plays an important role in learning and so investigating the role of affect or emotions during collaborative learning is an important area of research (Järvelä & Hadwin, 2013). Arousal and valence, which indicate affect (Russell, 2003), can be inferred from video data and used to evaluate collaborative learning (Dindar et al., 2020; Hayashi, 2019). For instance, Dindar et al. (2020) attempted to characterize collaboration quality by identifying leaders and followers in a collaborative task using the degree of emotional mimicry. Hayashi (2019) identified that the process of developing mutual understanding during a collaborative task is correlated with negative emotions. Additionally, the relationship between physiological synchrony and collaboration quality has been explored (Malmberg, Haataja, et al., 2019a; Pijeira-díaz et al., 2019; Schneider et al., 2020) and initial results suggest that physiological synchrony can be an indicator for collaboration quality. For instance, Schneider et al. (2020) used electrodermal data and identified a metric related to the number of cycles between low and high synchronization to be significantly correlated with collaboration quality and learning outcomes. Together this research suggests that the role of affective and physiological indicators on collaborative learning is still unclear and mediated by other factors. Hence it is important to look at these indicators along with other indicators such as speech and actions, while evaluating collaborative learning.

Building multi-modal models of collaborative learning

The literature above shows that several indicators impact collaborative learning, sometimes in contradictory ways. For instance, while some research suggests that transactivity in actions is not related to good collaboration (Popov et al., 2017), other research shows that transactivity in dialogue is indeed related to collaborative learning outcomes (Schneider & Pea, 2015). Another example is that while Popov et al. (2017) suggest that synchrony in actions is not related to good collaboration, other research suggests that synchrony in gaze is indicative of high quality of collaboration (Schneider & Pea, 2013). These complicated findings suggest that the effectiveness of collaborative learning in open-ended activities depends on multiple interconnected indicators. Recent research therefore investigates collaborative learning by combining multiple indicators obtained through multi-modal data sources in order to develop a richer and more comprehensive understanding of the learning mechanisms.

Empirical results suggest that combining multiple sources of data can provide better predictions of collaborative learning outcomes than any single modality of data alone (Emerson et al., 2020; Giannakos et al., 2019; Huang et al., 2019; Liu et al., 2018; Malmberg, Järvelä, et al., 2019b; Olsen et al., 2020; Spikol et al., 2018; Vrzakova et al., 2020; Worsley & Blikstein, 2018). Vrzakova et al. (2020), for instance, examined collaborative problem solving among triads and explored combinations of speech, actions and body posture patterns, which correlate with task performance. They found that certain multi-modal patterns are better than unimodal patterns for predicting performance. Olsen et al. (2020) investigated collaborative learning outcomes in an intelligent tutoring system and found

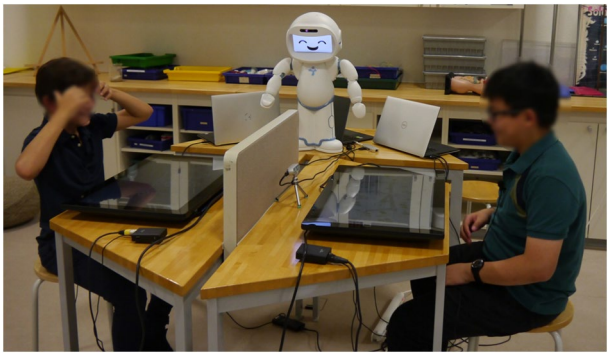


Fig. 1 JUSThink: a dyad interacts in a collaborative setup consisting of touch screens and a QTrobot

that combining modalities such as dual gaze, tutor log, audio and dialog provides more accurate prediction of learning gains than models using a single modality.

While multi-modal learning analytics explores different combinations of data streams along with various machine learning methods, what is not yet clear is how these combinations of indicators characterize collaborative learning. In order to develop a richer understanding of the collaborative learning processes, it is necessary to develop multi-modal learning profiles of groups of learners collaborating. Huang et al. (2019) did this by combining eye gaze, physiological sensor and motion sensing data, and identified three multi-modal states and the transitions between them, that are significantly correlated with task performance and learning gains. In this paper, we add to this line of research by proposing an approach to build multi-modal collaborative learning profiles of dyads as they work on an open-ended task around interactive tabletops with a robot mediator.

The study

We make use of the data from our previous study conducted with a robot-mediated collaborative learning activity called JUSThink (Nasir, Norman, Bruno, & Dillenbourg, 2020c). JUSThink is an evolving learning activity where learners interact with a collaborative learning platform consisting of two screens and a QTrobot acting as a guide and mediator (see Fig. 1). The learning goal of the activity is to impart conceptual knowledge about minimum-spanning-tree problems. A minimum spanning tree is a connected, undirected, edge-weighted graph that connects all the nodes together without any cycles and with the minimum possible total edge weight cost. We designed a scripted, collaborative problem-based learning activity for the learning goal of gaining conceptual understanding about minimum spanning trees (more details in (Nasir, Norman, Bruno, & Dillenbourg, 2020c)). Research shows that collaboration can help group members learn more than individual learning on high-complexity tasks such as those used in our problem-based learning environment (Kirschner et al., 2011). However, collaboration is not necessarily effective without support and collaborative learning scripts have been shown to provide good support for collaboration (Rummel & Spada, 2005). Hence dyadic collaboration structures are scripted into the learning design via features such as *mandatory turn-taking* and providing *partial information* (Nasir, Norman, Bruno, & Dillenbourg, 2020c).

A total of 96 children aged 9 to 12 years, organized in 48 teams, participated in this study (which were later reduced to 64 children, that is 32 teams, because of incomplete

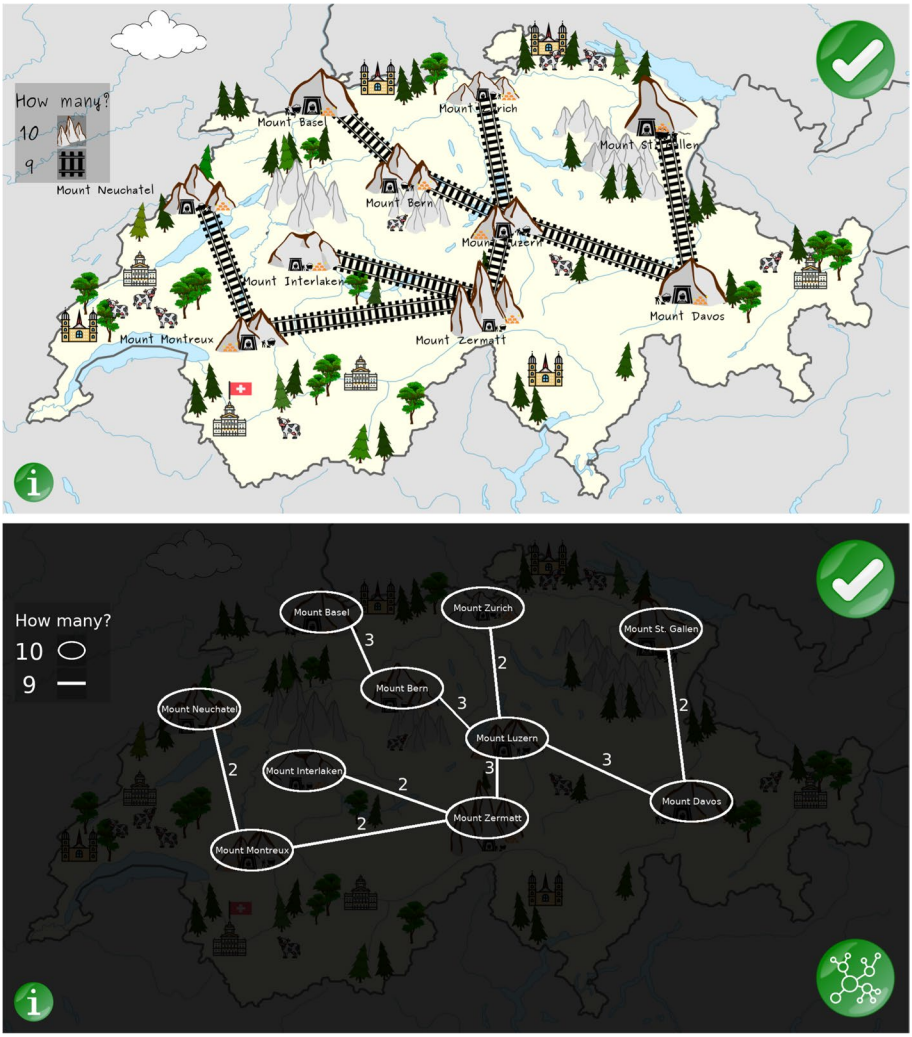


Fig. 2 The two views of the JUSThink game, namely *figurative* and *abstract*, as shown on the screens of the participants. The displayed set of tracks forms a minimum spanning tree to be constructed together by the participants

data due to real time technical issues either with sensors or logging). Each team spent a total of approximately 1 h for the entire activity that consisted of an introductory and briefing stage by the robot, a pre-test, the collaborative game-play, a post-test, a self-assessment questionnaire, and a goodbye phase. For the collaborative game play, the underlying concept of minimum-spanning tree is embedded in a scenario based on a map of Switzerland. To elaborate, there are fictional gold mines on the map depicted as mountains and labelled after Swiss cities, and the team has to help collect the gold by connecting these gold mines with railway tracks by spending as little money as possible. In graph terminology, each gold mine is a node, each railway track between two gold mines is an edge, and the money spent on each railway track is the cost of the edge. As shown in Fig. 2, there are two views in the game, namely *figurative* and *abstract*, and each learner in a team sees only one view

at a time. These views are swapped every two turns. In the *figurative* view, one can edit the graph by adding or removing tracks while in the *abstract* view, one can see the cost associated with building a track and can also open the team's previous solutions. Teams can submit solutions as many times as they want in the specified time period. They can also open a help page that describes game functionality and rules that the robot already elaborated for them before the game play.

For the data collection, we used one environment camera to capture the whole scene of interaction, two RGB-D front cameras, one for each child capturing the face up-close, and two lavalier microphones to capture audio data. The apps displayed on the touch screens and the robot communicated using the Robot Operating System (ROS) and the interaction on the touch screens (logs) was logged via ROSbag. The total number of teams considered for analysis was reduced from 48 to 32 due to technical and data completeness constraints (for more on the initial results, see (Nasir, Norman, Bruno, & Dillenbourg, 2020c)).

Methods

Dataset and preprocessing

As discussed above in the previous section, several learner measures can be used as indicators of collaborative learning. These can be divided into 1) behaviors, and 2) constructs. As mentioned earlier, a behavior is *an action or expression (verbal or facial) of the learner while interacting with the learning environment or a team member* that we extract from the log, audio, and video data streams of each participating dyad. These behaviors are representative of constructs that are *non-observable but have been linked to the process of learning*, such as *attention*, *exploration*, *reflection*, *frustration*, *confusion*, *excitement*, *synchrony* or *turn-taking* (Cherubini et al., 2008; Dindar et al., 2020; Hayashi, 2019; Martinez et al., 2011; Nasir et al., 2019; Sharma et al., 2015; Sharma et al., 2020; Weinberger & Fischer, 2006). As shown in Table 1, our dataset includes 28 multi-modal behaviors extracted from log, video and audio data, alongside performance metrics and various learning gains of the 32 teams. We use behaviors as indicators of the constructs based on similar work found in the literature. To begin with, the popular Russel's Core Affect Framework (Russell, 2003) states that an affect has a valence as well as an arousal component. Based on this widely adopted framework, negative valence and moderate to high levels of arousal are often linked with confusion and frustration, respectively, whereas positive valence and high arousal are indicative of excitement (Baker et al., 2010; Sharma et al., 2020). Inspired by this, we consider four features (*Positive Valence*, *Negative Valence*, *Difference in Valence* and *Arousal*) related to the emotional state of the team. The feature *Difference in Valence* is of interest as it immediately highlights that a team with a higher value has a positive emotional state. Please note that in this work, we do not distinguish between conceptual confusion and frustration as it is not straightforward to separate these accurately on the basis of the values of valence and arousal alone. For these reasons, we use the terms interchangeably when discussing our findings.

Similarly, gaze patterns have often been analyzed to gauge the attention of learners in collaborative settings (Schneider et al., 2016; Sharma et al., 2021). Therefore, here we extract the attention of the team to various parts of the screen, their partner, and the robot. Furthermore, in collaborative settings, speech measures have been widely used to measure the dynamics of the collaboration between the team members (Bassiou et al., 2016;

Table 1 Multi-modal features that represent behaviors and constructs

Construct	Marker	Behavior
Log Features		
Exploration	T_add	The number of times a team added an edge on the map
Exploration	T_ratio_add_rem	The ratio of addition of edges over deletion of edges by a team
Exploration	T_action	The total number of actions taken by a team (add, delete, submit, presses on the screen)
Exploration	Redundant_exist	The number of times they had redundant edges in their map
Reflection (Metacognition)	T remove	The number of times a team removed an edge from the map
Reflection (Metacognition)	T hist	The number of times a team opened the sub-window with history of their previous solutions
Reflection (Metacognition)	TI Tl add	The number of times a team, either member, followed the pattern consecutively: I delete, I add back
Reflection (Metacognition)	TI_Tl_delete	The number of times a team, either member, followed the pattern consecutively: I add, I then delete
Reflection (Metacognition)	Tl_T2_add	The number of times a team, either member, followed the pattern consecutively: I delete, You add back
Reflection (Metacognition)	Tl_T2_delete	The number of times a team, either member, followed the pattern consecutively: I add, You then delete
Usability Confusion	T help	The number of times a team opened the instructions manual
Video Features: Affective states and Gaze		
Emotional State	Positive Valence	The average value of negative valence for the team
Emotional State	Negative Valence	The average value of negative valence for the team
Emotional State	Difference in Valence	The difference of the average value of positive and negative valence for the team
Emotional State	Arousal	The average value of arousal for the team
Emotional State	Smile	The average percentage of time of a team smiling
Attention	Gaze at Partner	The average percentage of time a team has a team member looking at their partner
Attention	Gaze at Robot	The average percentage of time a team was looking at the robot
Attention	Gaze (Other)	The average percentage of time a team was looking in the direction opposite to the robot
Attention	Gaze at Screen_-Left	The average percentage of time a team was looking at the left side of the screen
Attention	Gaze at Screen_-Left	The average percentage of time a team was looking at the right side of the screen
Attention	Gaze Ratio of Screen_ Right and Screen_Left	The ratio of looking at the right side of the screen over the left side

Table 1 (continued)

Construct	Marker	Behavior
Audio Features: Speech		
Communication	Speech Activity	The average percentage of time a team was speaking over the entire duration of the task
Communication	Silence	The average percentage of time a team was silent over the entire duration of the task
Communication	Short Pauses	The average percentage of time a team pauses briefly (0.15 sec) over their speech activity
Communication	Long Pauses	The average percentage of time a team made long pauses (1.5 sec) over their entire speech activity
Communication	Speech Overlap	The average percentage of time the speech of the team members overlapped over the entire duration of the task
Communication	Overlap to Speech Ratio	The ratio of the speech overlap over the entire activity

Martinez et al., 2011; Viswanathan & VanLehn, 2017). We make use of several of these speech measures (*Speech Activity*, *Short Pauses*, *Long Pauses*, *Speech Overlap* and *Overlap_to_Speech_Ratio*) to capture talk time, and heterogeneity of verbal participation. The lengths of the *Short Pauses* and *Long Pauses* are based on findings from Campione and Véronis (2002) that, when analyzing pauses in various languages, found that pauses seem to support a categorization into brief (< 200 ms), medium (200-1000 ms), and long (>1000 ms) pauses. This is also echoed by the work of Heldner and Edlund (2010).

When it comes to interaction with a learning activity, log data such as frequency of actions has been used as an approximation for various constructs such as attention, engagement, interest, exploration, etc. (Martinez-Maldonado et al., 2013; Popov et al., 2017; Viswanathan & VanLehn, 2017). With our activity, we make use of frequency of actions such as additions, deletions, and redundant edges on the map (*T_add*, *T_remove*, *T_ratio_add_-del*, *Redundant_exist*). Furthermore, we are interested in actions or patterns that can indicate reflection. Consulting previously explored solutions is an indicator of reflection (Veenman, 2013). In addition to that, certain action patterns can also be indicative of reflection on self or partner's actions. Hence, we also consider such behaviors of looking at past solutions (*T_hist*), and correcting one's own or partner's actions on the go (*T1_T1_add*, *T1_T2_add*, *T1_T1_delete*, and *T1_T2_delete*) as indicators of reflection. Please note that we use *T_help* as an indicator of *usability confusion*, i.e., confusion with regard to the user interface and not as an indicator of conceptual confusion, which has been discussed previously.

In addition to these behaviors, we make use of one performance metric *last_error* that gives the error of the last submitted solution, where error can be defined as the difference between the cost of a submitted solution and the cost of an optimal/correct solution (optimal cost), normalized by the optimal cost (for an optimal solution, error will then be 0). We also use three types of learning gains *absolute*, *relative*, and *joint absolute* learning gains, which respectively measure how much the participant learned of all the knowledge available, how much the participant learned of the knowledge he/she did not possess before the activity, and the amount of knowledge acquired together by the team members during the activity. The team level values for the first two learning gains are calculated by taking the average of the individual learner values. It is important to mention that we distinguish between performance and learning such that performance measures the success/failure in the task itself via *last_error* whereas learning (*absolute*, *relative*, and *joint absolute*) measures the amount of knowledge gained during the interaction via a pre- and a post-test. Both of the tests are composed of 10 multiple-choice questions assessing concepts underlying the minimum spanning tree problem. The three types of learning gains are plotted versus the last error for all 32 teams in Fig. 3.

This dataset, with multi-modal behaviors as well as performance metrics and learning gains, has been made publicly available (Nasir, Bruno, et al., 2021a). It must be noted that, in our dataset, all behaviors are treated as cross-sectional averages and frequencies (i.e., over the entire duration of the task), and thus it is not time series data. The average value for the team for various behaviors is calculated by taking an average of the individual behaviors by each team member. Furthermore, for all the behaviors, data has been normalized across the teams, so that each behavior has a value between 0 and 1. This means that a value of 0 would be the lowest value of a behavior across all teams. Similarly, a value of 1 would be the highest value of a behavior across all teams. With respect to our previous work in Nasir, Bruno, et al. (2020; 2021), in this paper, we made slight changes to two of the behaviors *Short Pauses* and *Long Pauses* in the original dataset (Nasir, Norman, Bruno, Chetouani, & Dillenbourg, 2020b). Originally, the two pause behaviors were not

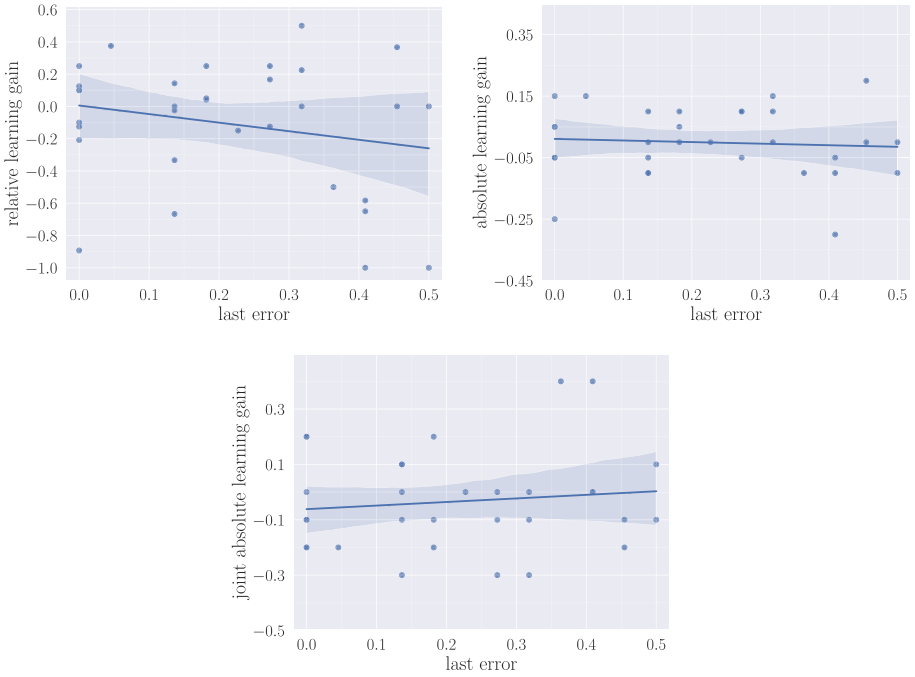


Fig. 3 Learning gains vs performance. All values here are non-normalized

normalized with respect to the teams' speech activity over the interaction. The change is motivated by the belief that normalizing the pause time gives a more accurate measure.

Lastly, we briefly elaborate on how the behaviors are operationalized. We extract log behaviors from the recorded rosbags while the behaviors related to both gaze and affective states are computed through the open source library OpenFace (Baltrušaitis et al., 2016), which returns both facial actions units (AUs) as well as gaze angles. In (Baltrušaitis et al., 2016), the authors validate their tool both in terms of AU recognition and eye gaze estimation among other features. For both, their tool performs better overall compared to other state-of-the-art methods. Facial Action Coding System (FACS), first presented by (Ekman & Friesen, 1978), is considered a major step in the research on facial expressions and is also considered to be the most widely used method for analyzing facial expressions (Cohn, 2006). The Facial Action Coding System made it possible to encode facial muscle movements, indicated by the AUs, to a corresponding displayed facial expression. A detailed table on each AU, its description, the facial muscle it corresponds to, and an example, can be found at IMotions blog[1]. The process of detecting AUs from human faces is now possible automatically with tools such as OpenFace, as first mentioned above. Certain combinations of these AUs can then be used to infer an emotional state[1] (Baltrusaitis et al., 2011; Benitez-Quiroz et al., 2016; El Kaliouby & Robinson, 2004). We use emotional states such as *valence* and *arousal*. Valence refers to the pleasantness and unpleasantness of an emotional stimulus (Kauschke et al., 2019). Further each emotional state is also linked to physiological arousal, such as anger and happiness being linked to increased autonomic response while sadness and boredom, are linked to decreased autonomic response (Herman et al., 2018). For generating quantitative values for positive and negative valence, we build on AUs that correspond to positive and negative emotions, respectively, based on the findings

Table 2 Action units employed for the calculation of positive and negative valence

Contructs	Action Units (AUs)	Corresponding Description	Corresponding Emotional States
Positive Valence	1,2,5,6,12,26	Inner Brow Raiser. Outer Brow Raiser. Upper Lid Raiser. Check Raiser, Lip Corner Puller. Jaw Drop	happiness, amusement surprised
Negative Valence	1,2,4,5,7,15,20,23,26	Inner Brow Raiser. Outer Brow Raiser. Brow Lowerer, Cheek Raiser. Lid Tightener. Lip Corner Depressor, Lip Stretcher. Lip Tightener, Jaw Drop	sad, angry, fear

from IMotions.[1] Authors in (Benitez-Quiroz et al., 2016) also conclude with similar findings. The AUs that we employ for positive and negative valence as well as their description and the emotional states they correspond to are shown in Table 2. After smoothing the data for each AU by employing exponential moving average, we take an average of the AUs to return the valence values. To calculate arousal, we use the average of all of those AUs listed in Table 2 for which the intensity is above a certain threshold. OpenFace not only returns the presence of an AU but also its intensity on a 5 point scale.

For voice activity detection (VAD), which classifies if a piece of audio is voiced or unvoiced, we made use of the python wrapper for the open source Google WebRTC VAD. WebRTC is a project that provides real-time communication capabilities for many different applications. This project is actively maintained by the Google WebRTC team[2] and due to it being open-source as well as reportedly one of the best and well maintained, there are several wrappers for it now, including for Python and Matlab. With the classification of voiced versus unvoiced frames for each student's audio channel, we can then generate all the features listed in Table 1.

Analysis approach

The goal of this work is to build and understand comprehensive multi-modal profiles of dyads who learn and those who don't as they work on JUSThink. To this end, we developed an analysis approach consisting of two parts: a quantitative approach and a qualitative approach. The quantitative approach relies on the outcomes of our previously published learning analytics technique (Nasir et al., 2021a, b; Nasir, Norman, Bruno, & Dillenbourg, 2020c) that helps identify groups of learners who have learning gains and those who don't. Using this approach, we are able to build the multimodal behavioral profiles for each group of learners. Since the profiles will be built based on our previous technique, we review it briefly in Multi-modal learning analytics, along with the method applied on the outcomes of this technique. The goal of the qualitative approach is to allow us to better interpret the multi-modal profiles and understand the learning mechanisms at play within each group of learners previously identified. We do this by interaction analysis of cases wherein we study

[1] https://imotions.com/blog/facial-action-coding-system/
[2] https://webrtc.org/

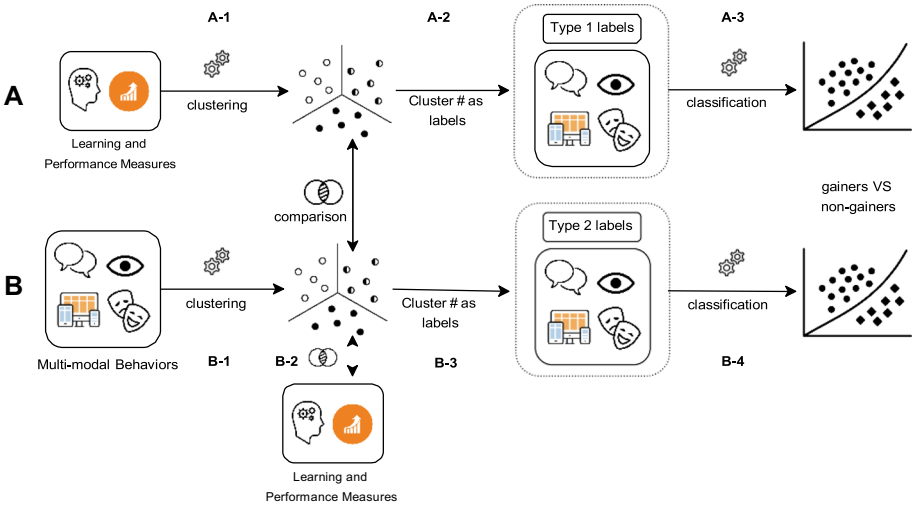


Fig. 4 Overview of our technique in Nasir, Norman, Bruno, and Dillenbourg (2020c)

the multi-modal behaviors from dyads within specific episodes of activity in each group of learners. We then unpack the likely multi-modal learning mechanisms at play. The choice of episodes will be based upon the findings of the quantitative approach; specifically we will focus on episodes where certain behaviors of interest identified in the quantitative approach are highlighted. Further details are in Interaction analysis of multi-modal cases.

Multi-modal learning analytics

The technique is visually presented in Fig. 4. It consists of two approaches: 1) *approach A*, which can be considered as a backward approach as it connects the learning outcomes back to the behaviors observed during the learning process and 2) *approach B*, that can be considered as a forward approach as it helps to move from multi-modal behaviors to learning outcomes. For the remainder of this paper, we will refer to them as *approach A* and *approach B*. This technique adopts a data-driven approach to identify labels linking behavioral profiles and learning in an effort to find *productively engaged* state of learners (Nasir et al., 2020c). It must be noted that in this paper, we use the term *gainers* in the context of the sub-group of learners who end up having learning gains, while the term *non-gainers* refers to those learners that do not have positive learning gains.

Approach A This approach starts with clustering on the learning as well as performance metrics of the teams as shown in step A-1. We then use these cluster labels as the ground truth for a classifier trained on multi-modal behaviors of the learners as shown in A-2 and A-3. This approach has been applied within learning analytics to identify the behavioral profiles of gainers vs non-gainers or high vs low performers (Kinnebrew et al., 2013; Worsley & Blikstein, 2011). In our case, the clustering reveals four clusters (more details on the four clusters in appendix A). Then, as a step towards building profiles, we perform a Kruskal-Wallis analysis on each pair of clusters to identify the significantly discriminating behaviors between each pair. However, we observe no significantly discriminating behavior between each pair. Such analysis on the four clusters from *approach A* can raise a

Table 3 The three clusters formed through Approach B, with mean values for learning gains (LG) as well as the last error specified. Statistical significance is indicated with *

Cluster name	Last_Error	Absolute_LG	Relative_LG	Joint_Absolute_LG	N
Approach B					
Expressive explorers	0.461	0.678*	0.693*	0.714*	14
Calm tinkerers	0.393	0.616	0.604	0.607	12
Silent wanderers	0.393	0.383*	0.348*	0.428*	6

misunderstanding that all learners, irrespective of learning or performance, exhibit similar multi-modal behaviors. It must be noted however, that this approach assumes *by design* that each of the learning and performance profiles (given by one cluster) is associated with a unique set of behaviors. However, what if teams with similar learning and performance actually exhibit two or more different sets of behaviors? This is the motivation for adopting *Approach B*, which represents a perspective shift in order to take such a possibility into account.

Approach B As depicted in step B-1 (Fig. 4), this approach begins with clustering the teams based on their multi-modal behaviors in order to identify the different behavioral profiles existing within the data. We then compare these behavioral clusters in terms of the learning gains and performance metric of the teams (B-2) in order to identify differences between the behavioral profiles in terms of their learning and performance. This is followed by comparing the clusters obtained in both approaches with respect to the teams they consist of. If, as we hypothesized, there are indeed multiple sets of behaviors associated with learning then 1) we should observe significant differences among some of the approach B clusters in regard to their learning gains (requirement 1) as well as 2) there should to be a one-to-many or many-to-many comprehensible mapping between the clusters from both approaches (requirement 2). This second requirement would mean that approach B provides us with distinct variants of behavioral profiles for the same learning profiles. To reiterate, while requirement 1 highlights that indeed gainers and non-gainers have different behaviors, requirement 2 is necessary to validate the existence of multiple behavioral profiles for the same type of learning profile, which is the motivation behind this approach. If the two requirements are met, the cluster labels from approach B can be employed as ground truth for a classifier as shown in steps B-3 and B-4. Thus this approach allows us to extract the differences that exist within both learning and behavior data, and align them to create multiple learning profiles.

The classification results have been published previously in Nasir, Bruno, and Dillenbourg (2020a) and reported here, with a minor change in the definitions of two behaviors, described in appendix A for ease of reference and transparency. As we obtained excellent classification results from *approach B*, in this paper, we focus in-depth on building behavioral profiles from the clusters resulting from this approach. As shown in Table 3, *Approach B* gives 3 behavioral clusters with the first two exhibiting high learning and the third lower learning; hence, with respect to learning, the groups can be named as *type 1 gainers*, *type 2 gainers*, and *non-gainers*, respectively. Note that the performance (*Last_Error*) in the task is very similar for each group. As done with the *Approach A* clusters, we now proceed to compare the resulting clusters in terms of their multi-modal behaviors by first performing an analysis of variance on the three clusters obtained using *Approach B* and then performing a Kruskal-Wallis analysis on each pair of clusters to identify the

behaviors that differ significantly between them. Indeed, we observe several discriminating behaviors between each pair, and so with respect to these behaviors, which will be seen in more detail in the upcoming sections, we name the groups of *type 1 gainers*, *type 2 gainers*, and *non-gainers* as *Expressive Explorers*, *Calm Tinkerers*, and *Silent Wanderers*, respectively. From this point on, we will use the two types of names interchangeably. Based on these quantitative findings, we then qualitatively analyze each group of learners as described in the following subsection.

Interaction analysis of multi-modal cases

In order to better interpret the multi-modal behavioral profiles identified above and attribute the likely learning mechanisms occurring in each group of learners, we qualitatively analyze a learning episode from each group. To do the analysis we select episodes when a "behavior of interest" is high. The exact behavior of interest was selected based on the results of the quantitative analysis, which we described in "Multi-modal Learning Analytics" section. The procedure makes sense for the purpose of unpacking a behavior that discriminates students who learned from those who did not. The selected behaviors were ones for which the effect of those behaviors on learning is not well understood. We analyze three episodes, targeting one randomly selected dyad from each of the three clusters. The random selection was hoped to maximize the chance that the selected dyad was representative of the cluster. As our quantitative results aggregate behaviours over the entire activity, these cases are meant to be illustrative of the likely underlying learning mechanisms during certain episodes when a behaviour of interest is high.

We begin by extracting the dialogue of the learners during the selected episode. The full transcripts can be found in the publicly available JUSThink dialogue and actions corpus by Norman et al. (2021). This corpus relies on manual transcription, due to the poor performance of state-of-the-art automatic speech recognition systems on this dataset, which consists of children's speech with music playing in the background. A graduate student completed two passes on each transcript, which were then checked by another native English speaking graduate student with experience in transcription/annotation tasks. We augment the dialogue transcript with average values of other behaviors during this episode to build a multi-modal transcript. We then interleave the dialogue, action and affective states to unpack how learning is happening within each episode. We perform interaction analysis of each episode with the analytic focus of turn-taking. The goal is to understand how turn-taking leads to learning during the episode, specifically the relationship between the content of the speech, the actions, the affect of the learners and their learning outcomes. Thus, with both the quantitative and qualitative methods aforementioned, we make an attempt to explain what learners' visible behaviors reveal about how learning happens in a collaborative constructivist learning activity.

Results

Pairwise significantly distinct behaviors

From Fig. 5, we observe that the behaviors with the highest variance among the three clusters come from all three modalities pertaining to log, speech and affective features.

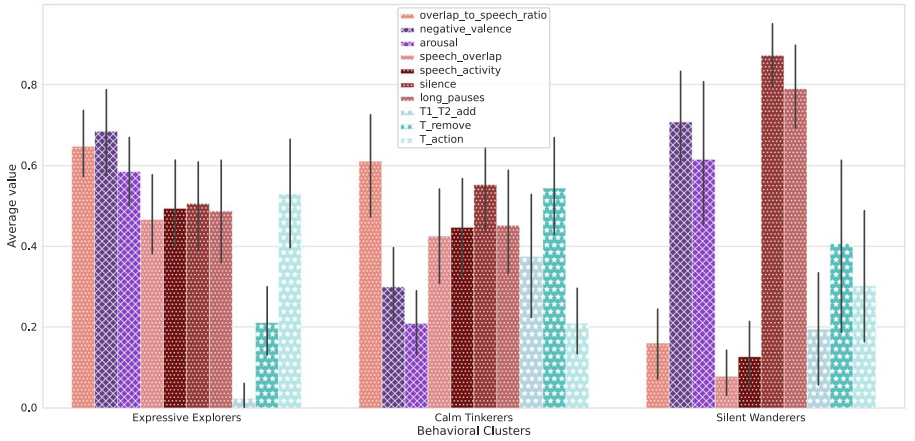


Fig. 5 Features with highest variance between all three behavioral clusters. For the ease of comprehension, each modality is represented by a unique pattern and each behavior within a modality by several shades of the same color

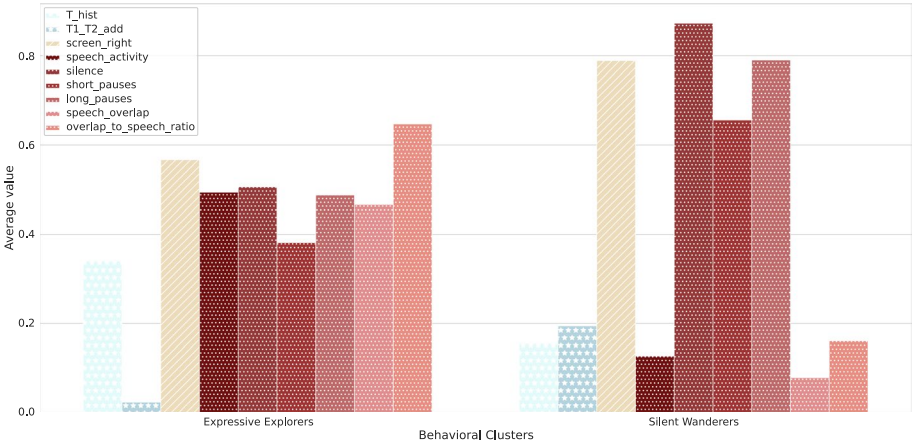


Fig. 6 Significant distinguishing features between the *Expressive Explorers* and the *Silent Wanderers*

Overall, it is clear that the manner in which each of the dyads interacted with the task (*T1_T2_add*, *T_remove*, *T_action*) is unique. Speech behavior (*speech_overlap*, *speech_activity*, *silence*, *overlap_to_speech_ratio*), on the other hand, is similar in the two gainer groups but very different from the *Silent Wanderers*. An interesting observation with regard to the affective features (*negative_valence*, *arousal*) is that the *Silent Wanderers* exhibit very similar arousal and negative valence behaviors to *Expressive Explorers* and both these groups differ from the *Calm Tinkerers* in this behavior. We elaborate on the differences between each pair of groups below. Note that in the upcoming figures, for the ease of comprehension, each modality is represented by a unique pattern, and each behavior within a modality is represented by several shades of the same color.

Table 4 p-values for the Kruskal-Wallis analysis on each pair with significance level of 0.05

Markers	Expressive Explorers and Silent Wanderers	Calm Tinkerers and Silent Wanderers	Expressive Explorers and Calm Tinkerers
Log Features			
T_add	0.14	0.85	0.03*
T_remove	0.13	0.42	0.00*
T_ratio_add_rem	0.05*	0.16	0.00*
T_action	0.07	0.37	0.00*
T_hist	0.04*	0.60	0.01*
T_help	0.45	0.14	0.46
T1_T1_remove	0.50	0.03*	0.00*
T1_T1_add	0.92	0.73	0.76
T1_T2_remove	0.80	0.39	0.04*
T1_T2_add	0.01*	0.19	0.00*
Redundant_exist	0.07	0.00*	0.83
Video Features: Affective states and Gaze			
Positive Valence	0.74	0.07	0.00*
Negative Valence	0.80	0.00*	0.00*
Difference in Valence	0.62	0.22	0.16
Arousal	0.93	0.00*	0.00*
Smile	0.93	0.05*	0.01*
Gaze at Partner	0.28	0.45	0.12
Gaze at Screen_Left	0.11	0.01*	0.23
Gaze at Screen_Right	0.02*	0.22	0.53
Gaze Ratio of Screen_-Right and Screen_Left	0.28	0.45	0.83
Gaze at Robot	0.50	0.22	0.16
Gaze (Other)	0.45	0.16	0.04*
Audio Features: Speech			
Speech Activity	0.00*	0.00*	0.71
Silence	0.00*	0.00*	0.71
Short Pauses	0.04*	0.16	0.23
Long Pauses	0.01*	0.01*	0.60
Speech Overlap	0.00*	0.00*	0.68
Overlap to Speech Ratio	0.00*	0.00*	1.00

Expressive explorers and silent wanderers

Figure 6 shows the features that are significantly different between *Expressive Explorers* and *Silent Wanderers*. The corresponding p-values are listed in Table 4. Concerning log features, we observe that *Expressive Explorers*, relative to *Silent Wanderers*, do significantly fewer actions of the sort where one team member deletes an edge and the other adds it back (*T1_T2_add*). At the same time, they look at their previous solutions (*T_hist*) significantly more than *Silent Wanderers*. This suggests that *Expressive Explorers* perform more

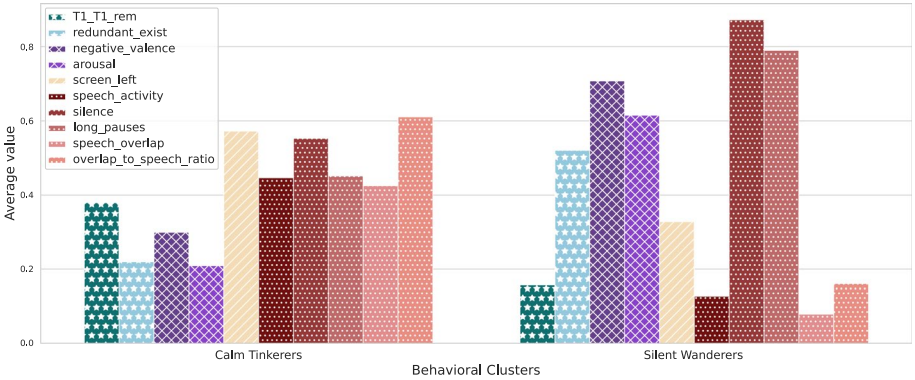


Fig. 7 Significant distinguishing features between the *Calm Tinkerers* and the *Silent Wanderers*

global reflection, i.e., reflection on their previously constructed solutions, while *Silent Wanderers* do more local reflection, i.e., reflection on their most recent actions.

Apart from this difference, the two groups are also significantly different in their speech behavior. *Expressive Explorers* not only speak more between themselves (*Speech Activity*), but also have lower number of short and long pauses (*Short Pauses*, *Long Pauses*) when they speak and a higher degree of overlap (*Speech Overlap*, *Overlap_to_Speech_Ratio*) when interacting. Finally, the two groups show no significant difference in their affective features, as seen by the fact that both *Expressive Explorers* and *Silent Wanderers* displayed very similar valence and arousal behaviors (specifically high *arousal* and high *negative valence*).

Calm tinkerers and silent wanderers

Looking at the distinguishing behaviors between *Calm Tinkerers* and *Silent Wanderers* (see Fig. 7 and Table 4 for the p-values of the KW tests), we observe that the differences lie in the way in which they interact with the task itself, their speech behavior and also their affective features. Unlike *Expressive Explorers*, the *Calm Tinkerers*, relative to *Silent Wanderers*, do more of local reflective actions, where a team member adds an edge and then removes it right after (*T1_T1_rem*). Moreover, while *Calm Tinkerers* carefully minimize the number of redundant edges (i.e., two alternative paths connecting location A with location B) present at any time on their map in the task, *Silent Wanderers* allow for such redundancies to be present on the map significantly more.

In terms of their speech behavior, *Calm Tinkerers* have higher speech activity (*Speech Activity*), lower number of long pauses (*Long Pauses*) and higher speech overlap (*Speech Overlap*, *Overlap_to_Speech_Ratio*) than *Silent Wanderers* (who are non-gainers in terms of learning). It is important to remark that the same difference was observed between *Expressive Explorers* and *Silent Wanderers*, thus suggesting that speech behaviors exhibit some generality across behavior profiles in distinguishing gainers from non-gainers. Lastly, this group of gainers displays significantly lower negative valence and arousal (*negative_valence*, *arousal*) compared to the *Silent Wanderers*, illustrating one way that *Calm Tinkerers appear* relatively calmer.

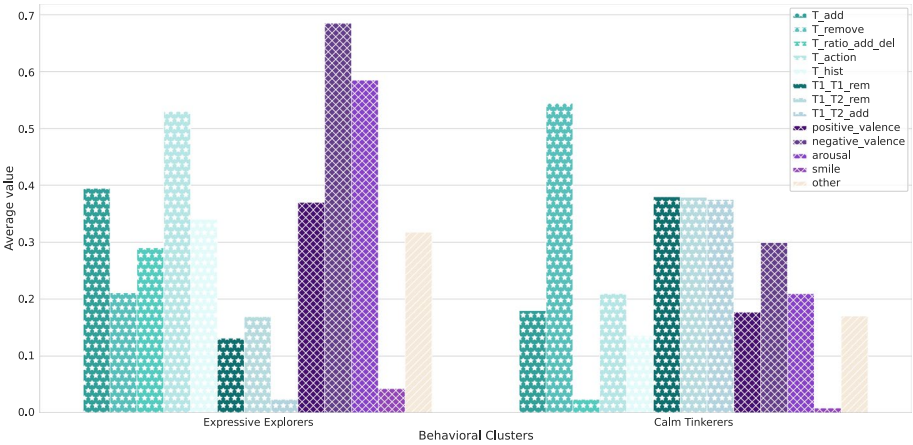


Fig. 8 Significant distinguishing features between the two type of gainers

Expressive explorers and calm tinkerers

Lastly, we compare the significant distinguishing behaviors between two types of gainers (see Fig. 8 and Table 4 for the p-values of the KW tests). We observe that the two groups of gainers significantly differ in most of their log behaviors. If we look closely at these behaviors, we observe that *Expressive Explorers* do more actions (*T_action*) in general, specifically doing more edge additions (*T_add*) and, consequently, displaying a higher ratio of adding to deleting edges (*T_ratio_add_rem*). Furthermore, they open their history significantly more times (*T_hist*). *Calm Tinkerers*, on the other hand, have more deletion actions (*T_remove*) and a higher number of addition-deletion action sequences of the type *T1_T1_-rem*, *T1_T2_rem* and *T1_T2_add*. These findings suggest that *Expressive Explorers* enact a global exploratory approach characterized by global reflection on previous solutions while *Calm Tinkerers* exhibit a local exploratory approach where they carry out in-the-moment reflection and correct their own and their partners' actions on the go, which can be described as local reflection. For example, *Expressive Explorers* successively add edges on the map and then look at the cost effectiveness of their constructed map by comparing it with their past solutions, while *Calm Tinkerers* show a pattern of adding an edge and then deleting it right after or vice versa which may be triggered due to reflection. A specific example will follow in the case studies discussed in "Interaction Analysis of Multi-Modal Cases" section.

Moreover, *Expressive Explorers* have higher average values of *valence* and *arousal* compared to the *Calm Tinkerers*, suggesting that they were more expressive in their interactions. These results show that gainers can exhibit a profile associated either with frustration or calmness. Lastly, notice how none of the speech behaviors differ significantly between the two types of gainers, once again pointing to the fact that gainers, irrespective of their other behaviors, all had a similar speech behavior quantitatively.

Interaction analysis of multi-modal cases

As shown above, *speech overlap* is a behavior that distinguishes *Silent Wanderers* (who do not overlap with one another as much) from both types of gainers (who overlap significantly more). Specifically, our results suggest that a high amount of overlapping speech can be more productive for learning relative to when there is less speech overlap. It has been reported in the literature (Bassiou et al., 2016) that speech overlap is one of the speech features that distinguishes the quality of collaboration. However, this literature also suggests that the frequency of overlaps is negatively correlated with collaboration in children (Kim et al., 2015). Given these contradictory findings on the role of overlapping speech in collaborative learning, we consider "speech overlap" as a behavior of interest for qualitative analysis. We seek to understand the nature of overlapping speech and turn-taking during the task. Specifically, for one randomly selected team from each group of learners, we pick a chunk of dialogue of a few seconds, that corresponds to the first time a team reaches the highest level of speech overlap to speech ratio (*overlap_to_speech_ratio*) consecutively for the whole duration of the chunk. We report below the dialogues taking place between the team members, along with the averages of their actions and affect within this duration. The blue and red colored rectangles in the upcoming figures highlighting dialogue indicate the duration in which learner A and B are speaking, respectively; hence, highlighting speech overlap when the rectangles overlap. The start and the end time for the dialogues (in seconds) are also indicated in the figures. Right next to the dialogues, in these figures, we also report other behaviors for each chunk. Our temporal data for this qualitative analysis is organized in 10 s windows. We use the values in these windows to report both the average of these behaviors over the *entire interaction* and *within the chosen chunks* (that range from 30 to 60 s), one for each team in the case studies. Note that we do not include gaze behaviors as gaze was not found to be a significant distinguishing behavior in our quantitative analysis. Lastly, Fig. 9 shows the state of the two views, i.e. an empty map on which the learners build a solution together, that we will be referring to in the next sections.

Episode from expressive explorers

The dialogue excerpt, shown in Fig. 10, occurs right after the dyad submitted a solution and was informed by the robot that it is not the optimal solution yet. Hence, what the participants see on their screens at the time when this dialogue starts is an empty map, as shown in Fig. 9, i.e., a map that has been cleared after a solution was submitted. The team can now start building a new solution on this empty map.

We observe that both team members interject when the other is speaking. However the content of the contributions builds on the partner's conceptual ideas, which is conducive to the emergence of novel, integrative solution ideas. The high speech overlap is thus not caused by a lack of collaboration but a high degree of understanding between the team members, owing to which they are "completing each other's sentences". In addition, we observe that the average values of arousal and negative valence during this exchange are lower (0.22 and 0.20 respectively) than the average values (0.34 and 0.28) of this team over the entire task, suggesting a shift towards low arousal states such as "neutral", "boredom" or even "sadness" right after one hears feedback on their solution. This is interesting because *Expressive Explorers* exhibit a higher level of frustration overall.

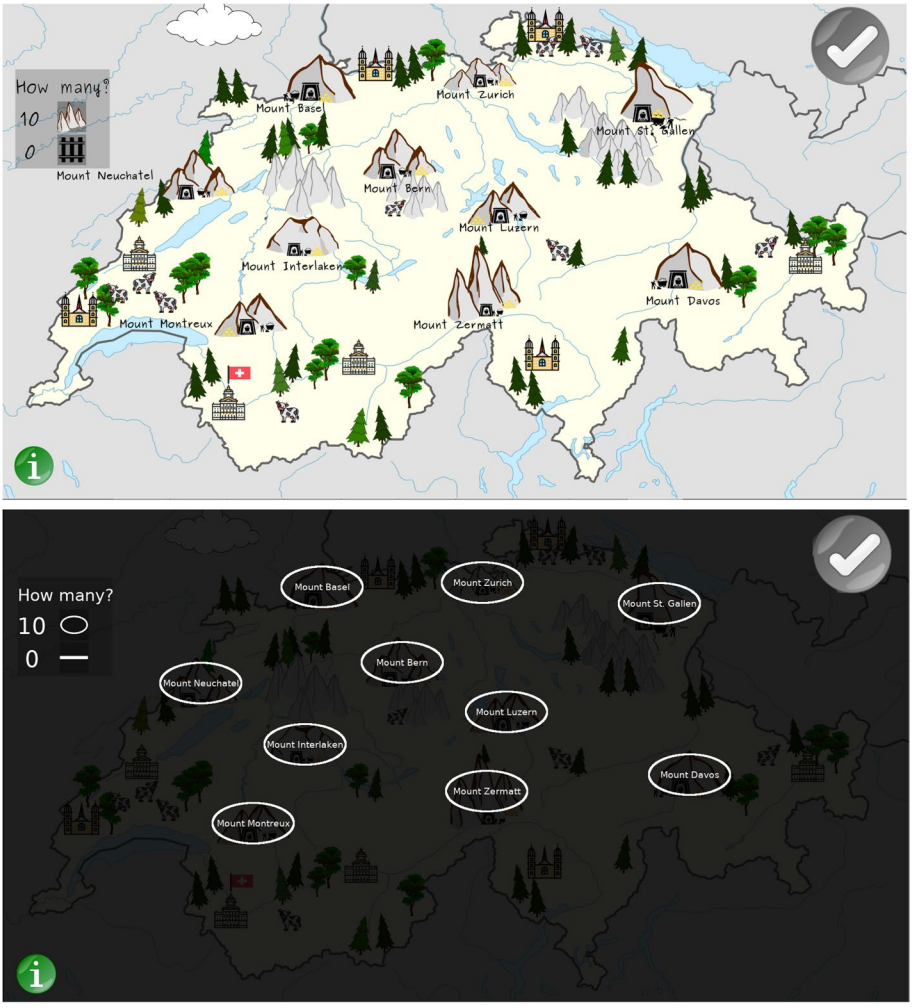
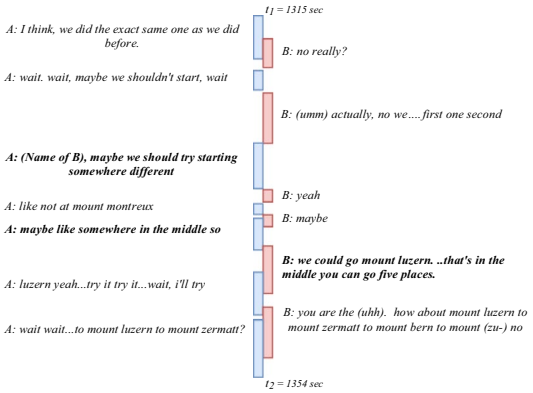


Fig. 9 The two views of the JUSThink game, namely *figurative* and *abstract*, as shown on the screens of the participants when they are empty

Now looking at the log actions of this team during this chunk with respect to the whole task, we observe that the team employs a more global exploration strategy with an increase in both addition actions and reflection in terms of looking more at their history. As seen in the bold section of the dialogue above, the team reassesses the foundations of their approach and then revises it. Further, looking at the ratio between additions and deleted actions in this chunk versus over the whole task, we note that the team is only doing additions. This may be because in this time the team is starting from an empty map and building a new solution. Connecting these observations back to the overall solution strategies of these types of gainers, this episode provides deeper multimodal insights for how this type of gainers learn through a more global exploratory approach and reflection on their overall solution strategy.



	Average over entire interaction	Average over the chunk
Arousal	0.34	0.22
Negative Valence	0.28	0.20
Addition	0.22	0.4
Deletions	0.13	0
History	0.09	0.12
Local Reflective actions	0	0

Fig. 10 The dialogue for an *Expressive Explorers* team where the blue and red rectangles indicate the duration in which learner A and B are speaking, respectively. Speech overlap is indicated by the overlapping rectangles. Other relevant log and affective features are also shown in a parallel table displayed on the right

Episode from calm tinkerers

The two views on the respective screens of a random team of *Calm Tinkerers*, at the time the relevant dialogue starts, are as shown in Fig. 11. This snippet of dialogue, shown in Fig. 12, occurs approximately one minute after they submitted their first solution and were told it is not the optimal solution yet.

In this excerpt, the team members are attempting to optimize the solution by adding a particular edge ("Bern to Interlaken") to the solution. Firstly, when both team members agree upon the overall strategy, they both speak over each other to complete the steps to be taken towards the solution. Secondly, when there is disagreement about the next action, there is a high overlap of speech; however the dialogue leads to an agreement on the action to be taken. Thus, the high degree of overlap seems to be related to these cycles of proposal-negation-agreement, which could be one mechanism by which the locally reflective problem solving strategy is manifested in this group of learners. Indeed, as the dialogue shows, the team members immediately reflect and correct each other's actions. This is a sign of negotiation that is inherent in a collaborative problem solving session and that leads to mutual understanding of the solution space.

Zooming into the teams' affective state during this exchange, we find that the average arousal and negative valence in this chunk was 0.38 and 0.30 respectively, which is higher than the team's average arousal and negative valence (0.32 and 0.23 respectively) over the entirety of the interaction. This indicates that during this period of high speech overlap, the team was in a higher state of arousal, which could possibly indicate a state of disequilibrium as suggested by previous research (D'Mello & Graesser, 2012; Lodge et al., 2018). Recall that *Calm Tinkerers* overall exhibit lesser frustration than the other two types of learners.

In terms of actions, we see that in this chunk the teams' ratio of deletions to additions is higher than over their entire interaction. This could be because by this time the team had already added several edges towards a potential solution and were deleting edges through the negotiation and optimization process seen in the dialogue above. Further we see that none of the other actions signifying reflection, such as looking at their history or deleting their own or their partners edges is seen here. Connecting back to *Calm Tinkerers'* overall

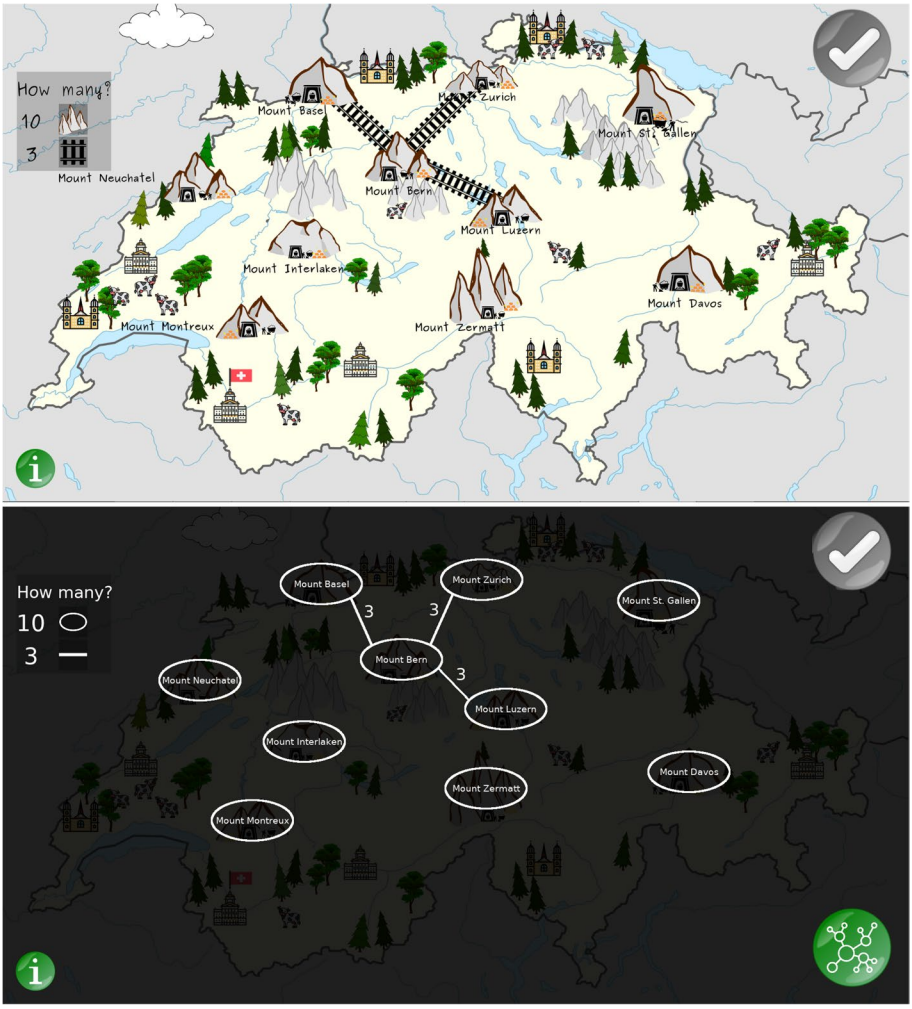
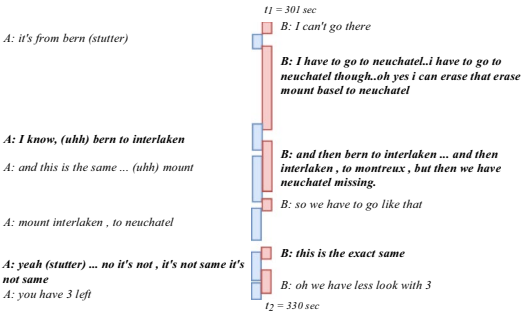


Fig. 11 The two views of the JUSThink game, namely *figurative* and *abstract*, as shown on the screens of the participants, in this case, from a team belonging to the group of *Calm Tinkerers*



	Average over entire interaction	Average over the chunk
Arousal	0.32	0.38
Negative Valence	0.23	0.30
Addition	0.31	0.25
Deletions	0.12	0.12
History	0.10	0
Local Reflective actions	0.75	0

Fig. 12 The dialogue for a *Calm Tinkerers* team

	Average over entire interaction	Average over the chunk
Arousal	0.51	0.50
Negative Valence	0.37	0.40
Addition	0.24	0.2
Deletions	0.12	0
History	0.07	0.07
Local Reflective actions	0	0

Fig. 13 The dialogue for a non-gainer team of *Silent Wanderers*

solution strategy, we see that this chunk demonstrates how these teams learn through a local exploration strategy of additions and deletions, rather than reflecting on overall strategy by looking at their history.

Episode from silent wanderers

This dialogue, shown in Fig. 13, takes place within a team of *Silent Wanderers* right after they submitted a solution and were told by the robot that it is not optimal. Hence, when the dialogue starts, the screens for this team also shows empty maps, as in Fig. 9. At that point, the dyad was able to start building a new solution.

In this dialogue, we observe that the team first agrees on the goal to achieve (a solution cost of 24). However, the initial idea put forth by a team member (B) is not taken up by A, which leads to a cycle of proposal-negation-agreement. The negotiation between the team takes longer compared to *Calm Tinkerers,* but they eventually come to an agreement about the first action to take while building a new solution (connecting "Basel to Zurich"). During this negotiation the team members speak over each other, as also seen with the *Calm Tinkerers,* which indicates constructive collaboration since this non-gainer team also reaches an agreement on a path forward to the solution. Overall, however, as seen from our quantitative analysis, the duration of such speech overlap is significantly less in the *Silent Wanderers*.

It is interesting to note that the arousal of this non-gainer team sees an increase followed by a dip (ranging from 0.57 to 0.45) during this exchange, compared to the teams' average arousal of 0.51. The dip in arousal during this exchange that occurs right after getting the feedback on their solution (which is very far from optimal) suggests a tendency towards low arousal emotions such as "neutral", "sadness" or "boredom". On average, however, this team's arousal and negative valence over this chunk (0.5 and 0.4 respectively) is similar to

their arousal and negative valence over the entire task (0.51 and 0.37 respectively). We recall that non-gainer teams on average exhibit higher frustration than the gainer teams.

In terms of actions, we observe that in this chunk, where they begin from an empty map, the team performs only additions and no deletions as they try to negotiate and build a better solution. Further, their reflective actions (looking at history and deleting their own or their partner's actions) are similar to their reflective actions across the entire task. Recall that non-gainer teams on average do fewer reflective actions of any type.

To summarize, while the non-gainer team, similar to the two gainer teams, similarly exhibits constructive communication during an episode of high speech overlap, they do not demonstrate any change in their reflective actions during this chunk right after a "failure" or exhibit any change in their affective state. Further, they have significantly shorter durations of such speech overlap over the entire task as compared to both types of gainers. This, along with the fact that they have fewer reflective actions overall, could be a reason for their learning process not being as effective as the gainers.

Discussion

The goal of this paper is to build a multi-modal understanding of learning vs non-learning as it happens in a collaborative open-ended activity. Our combined multi-modal learning analytics and interaction analysis methodology enabled us to identify two multi-modal profiles of learners who have learning gains and one multi-modal profile of learners who do not have learning gains. Now that we have quantitatively compared the profiles pair-wise in "Pairwise Significantly Distinct Behaviors" section and qualitatively compared three teams, one from each profile, separately in "Interaction Analysis of Multi-Modal Cases" section, in this section, we begin by discussing each of the three profiles of learners with respect to each modality. Next, we discuss how multi-modality furthers our understanding of collaborative learning and how the outcomes might contribute to designing effective interventions in similar CSCL settings.

Speech activity

In terms of speech behaviors, such as the amount of speech activity of a team, both types of gainers exhibit a very similar behavior quantitatively, that is significantly different from the one displayed by the *Silent Wanderers*. The same is true for other speech behaviors including speech overlap between team members, the overlap to speech activity ratio, and short and long pauses over the entire speech activity. Overall, we find that there is a lot more verbal interaction within the teams that complete the activity with higher learning gains, as observed in previous research on collaborative learning (Bassiou et al., 2016; Praharaj et al., 2021; Weinberger & Fischer, 2006). This is not surprising since the nature of the collaborative activity requires the learners to communicate, share information, and build common ground to enable construction of a solution (Barron, 2003; Roschelle & Teasley, 1995). As we highlighted in the episodes of high speech overlap dialogue, we observe two mechanisms of verbal interaction that support collaborative learning. In one case, the dyad demonstrates a high degree of transactivity, which is known to be good for learning (Teasley, 1997). This is seen by completion of each others' sentences as the speech overlap is a way to align on their plan for solution building. In the other two cases, we observe proposal-negation-agreement cycles (Barron, 2003; Roschelle, 1992) in the team members'

dialogue during these periods of high speech overlap, indicating that the process of proposal discussion and uptake was happening, which is also indicative of good collaboration (Barron, 2003). Hence, contrary to the literature that suggests that the frequency of overlaps is negatively correlated with collaboration in children (Kim et al., 2015), speech overlap seems to be an indicator of the negotiation that is inherent in the collaborative learning process as also found by Bassiou et al. (2016) and Praharaj et al. (2021). The difference in the learning of the *Silent Wanderers* could be because of fewer such productive collaborative episodes within this group. Lastly, both types of gainers show a significantly smaller percentage of long pauses in their speech relative to *Silent Wanderers,* which again, as suggested by previous research (Fors, 2015), tends to be indicative of better communication, which is essential for good collaboration.

Log actions

In terms of actions, it is clear that the two types of gainers do not exhibit the same exploratory approach, with *Expressive Explorers* showcasing a more global exploratory approach of building a solution, testing and reflecting on their previous solutions, and then building a new one, contrasting with *Calm Tinkerers* displaying a more local exploratory approach of adding edges, reflecting on and possibly deleting an edge in-the-moment, as they build the solution. On the other hand, *Silent Wanderers* seem to not be adhering strictly to either of the two strategies and rather are displaying a mix of both. However, as we observe, both approaches incorporate some form of reflection, that is generally less representative of the non-learning group both in terms of reflection-in-the moment and reflection-on-prior actions. This may be why there are more redundancies present on the map for them at any given point in time. Hence, in terms of interaction with the task, it is the act of regulating their solution-building approach through reflection that differentiates the gainers from the *Silent Wanderers*. This is not surprising since reflection has been found to play a pivotal role in learning in problem-based learning environments (Barron et al., 1998; Do-lenh, 2012; Etkina et al., 2010; Hmelo-Silver, 2004).

As suggested in prior research, regulating ones' own and a partners' cognition, metacognition, behaviors and emotions is important for productive collaborative learning (Järvelä et al., 2016). Our findings related to speech and actions together suggest that the gainers regulated their learning by verbally interacting with each other and reflecting on their solution approach, thus obtaining learning gains. What might explain why the *Silent Wanderers*, on the other hand, did not gain is that they had less verbal interaction and reflection.

Affective behaviors

When it comes to affective behaviors, we observe that *Expressive Explorers* exhibit high arousal and negative valence (possible confusion/frustration) similar to the non-learning group, *Silent Wanderers,* and significantly different from the second group of gainers, *Calm Tinkerers*. This suggests that confusion/frustration itself may not be the reason for them not learning and that it is rather the set of other behaviors that accompany this frustration that determine whether a team learns or not in an open-ended collaborative activity. This outcome is contrary to the more popular approach that treats frustration as something to alleviate (D'Mello & Graesser, 2012; Hone, 2006; Klein et al., 2002) but rather is in line with the work of Baker et al. (2010) and Mentis et al. (2007) that have suggested that in some cases, frustration may not need remediation. However, an important question that arises

here is whether the *Expressive Explorers* end up learning *despite* frustration or *because* of it. The answer to this question is out of the scope of this paper due to the correlational approach and small sample size; however, it can be an interesting question to explore for the community as well as for us in future work.

As highlighted by the interaction analysis, both types of gainers show a change in their average emotional state right after submitting a sub-optimal solution, together with a phase of high speech overlap. The team of *Expressive Explorers* show a dip in their emotional state while the team of *Calm Tinkerers* show an increase in their emotional state. The latter case can be explained by the model proposed by D'Mello and Graesser (2012) for the dynamics of affective states during complex learning, where the authors suggest that learners' states oscillate between a state of equilibrium (flow) and disequilibrium (confusion) when an impasse is detected. In the episode we analyzed, as the *Calm Tinkerers* discovered that their solution was incorrect, they were observed to become more confused (higher emotional states). The case of the *Expressive Explorers* team is interesting because it is not directly explained by the model of D'Mello and Graesser (2012). However, it must also be noted that this team in general showed higher frustration during the activity and thus this can be considered *their* state of equilibrium. Hence, on receiving feedback about the sub-optimality of their solution, they switched to a lower emotional state, which for them is a state of disequilibrium. In the case of both types of gainers, however, we see an attempt to regulate the state of disequilibrium via effortful reasoning and problem solving (Järvelä et al., 2016). This leads to an increase in verbal interaction with interjections while discussing revised problem solving strategies. It is interesting that the *Silent Wanderers* team showed no change in their affective state in the episode of high speech occurring after submitting a sub-optimal solution. It is worth exploring further what this lack of change in affective state at a moment of impasse means for learning.

Gaze behaviors

When it comes to gaze patterns, we did not observe any significant differences between the two gainer groups, suggesting that they exhibit very similar behavior when paying attention to the screen as well as when looking at their partner or the robot. Moreover, when comparing the two types of gainers with the *Silent Wanderers*, the only significant difference observed was with respect to looking more on the right (where the previous solutions can be displayed upon clicking on a button) or the left side of the screen, while there are no differences among the gaze patterns when looking towards their partner, the robot or the opposite side of the robot. This suggests that, for the gaze behaviors we considered, a "productive" gaze pattern does not emerge from the data.

Tying it all together: How do the different modalities interact?

Going back to our research question on multi-modal behavioral profiles of learning in a collaborative constructivist activity, we have identified two types of gainer profiles based on our pair-wise analysis in "Results" section. The first gainer profile, *Expressive Explorers,* consists of *effective communication* as seen by their high amount of verbal interaction between the team members, periods of high overlap in speech of the team members, fewer longer pauses in the speech; a *global exploratory approach* consisting of adding a lot more edges while solving the task, followed by *reflection* by opening their past solutions; and exhibiting a state of *frustration* seen by higher arousal and negative valence. The second

gainer profile, *Calm Tinkerers*, similar to the first one, is characterized by *effective communication*. However, differently from the first one, it consists of a *local exploration approach* in which team members remove a lot more edges while constructing a solution; *local reflection* or reflection-in-the moment, represented by a higher number of sequence actions such as a team member adding or removing their own or their partners' recently added edge; and a *relatively calm emotional state* characterized by lower arousal and negative valence. Finally, the non-gainer profile, i.e., that of *Silent Wanderers*, is characterized by *poorer communication,* meaning significantly less verbal interaction and less speech overlap, and more long pauses compared to the two types of gainer profiles. In addition, similar to *Expressive Explorers*, *Silent Wanderers* exhibit *frustration*; however, compared to both the gainer profiles, they *reflect* less both on prior solutions (open their history less) and recent actions (have fewer sequence actions such as a team member adding or removing their own or their partner's actions). This third profile lends further support for the need of regulation of learners' problem solving strategies and frustration via reflection and verbal communication in order for effective collaborative learning to happen (Järvelä et al., 2016).

The fact that only two out of three identified multi-modal behavioral profiles was associated with learning is in line with the literature suggesting that while collaboration can scaffold learning, it is contingent upon the quality of the interactions (Dillenbourg et al., 2009), and diverse and complex conditions (Lou et al., 2001; Meier et al., 2007). Furthermore, and consistent with the literature, we found that while impasses and failures can offer the conditions for learning to happen, whether it actually does happen depends on learners' cognitive (Barron, 2003; Lodge et al., 2018; Loibl et al., 2017), social (Weinberger & Fischer, 2006) and emotional behaviors (D'Mello & Graesser, 2012) as a response to the moment of encountering an impasse. Our work identifies two possible collections of actions, speech and affective behaviors associated with episodes where effective collaborative impasse-driven learning was observed to occur and one collection of behaviors where it was not. Thus, through this paper, we provide a more holistic assessment of the behaviors underlying collaborative impasse-driven learning that might contribute to refining the theories of both collaborative learning and impasse-driven learning as we elaborate below.

Our findings confirm some of the findings in the CSCL literature in the context of an open ended collaborative activity: (1) *verbal interaction in a constructivist collaborative activity, not just in terms of amount of speech but also overlap of speech between team members, serves as a discriminatory factor between gainers and non-gainers,* but (2) *it is not necessarily individual behaviors that discriminate gainers from non-gainers; rather it is a set of behaviors* which may not always be obvious when observed by experts such as a teacher or observer in such exploratory collaborative activities. Furthermore, it must also be noted that half of the gainers in *Expressive Explorers* and *Calm Tinkerers* groups actually *failed* at the task, and the same ratio holds in the *Silent Wanderers* group, suggesting once more that (3) *performance in the task, which often influences human experts in their evaluation of a learner's progress, is not always a reliable predictor of learning.*

What is relatively less clear from past literature is when high and low reflection or emotions are productive for learning. Our work takes a step in that direction, as the aggregate multimodal behavioral profiles of learners highlight that certain kinds of reflection (reflection-in-the-moment) is accompanied by calmer emotions, while other kinds of reflection (reflection-on-prior actions) is accompanied by more expressive emotions. That is, in our work, we discover that there exists a *relationship* between two of the modalities, i.e., *problem-solving strategy* and *emotional expressivity*, that can discriminate multiple ways of achieving the learning goal. The fact that the strategies differ among the two types of gainers is not a surprise as problem-solving strategies have been studied in CSCL literature;

		Emotions	
		Expressive	Calm
Strategies	Explorers	X	
	Tinkerers		X

Fig. 14 The interplay between the problem-solving strategies and the emotional expressivity for the gainer teams

however, the fact that the arousal and valence interact with the different types of strategies is a novel contribution of this work. More specifically, we observe the interplay in the diagonal shown in Fig. 14, that suggests that *expressivity of emotions is related to the problem-solving strategy*. A certain strategy leads to more episodes of frustration than the other, and examining multiple modalities simultaneously allows us to unearth this relationship. It also raises an interesting question as to why there are no gainer teams in the cross diagonal. Is this where the non-gainers lie? While the *Silent Wanderers* do exhibit a higher emotional expressivity, they do not strictly adhere to either of these two problem-solving strategies as they exhibit lower levels of both local and global reflection. Hence, they too do not lie in this cross diagonal. Then, the question to consider is whether the cross-diagonal would be associated with learning or non-learning profiles.

Hence, we argue that the insights from our current results might inform CSCL designers regarding *what* interplay between problem solving strategies and emotional expressivity may be more conducive to learning in such a CSCL setup in *addition* to the more obvious behavior of speech activity. This can help in making a more informed design of a robot or an autonomous agent for adaptive interventions, which might first use simple speech activity measures to identify non-gainers. Other speech measures such as semantics of speech, that might be more descriptive, need manual work by humans that cannot always be done in real time. Hence, the easier automatic assessment in real-time with speech activity measures makes them a great choice for guiding effective interventions by intelligent systems. Once an 'unproductive state' is identified via speech, the agent/robot can use information provided by the other modalities in an attempt to scaffold the learners towards either of the gainer profiles. For example, if a team is following a more tinkering problem-solving strategy and they continuously start displaying higher levels of frustration on average, there may be a need to remediate this frustration, as it could push them into a behavior pattern associated with non-gainers. Conversely, frustration displayed by a team displaying a more global exploratory problem strategy may not need remediation. While current learning profiles tie back to literature both in terms of behaviors and constructs as we see above, the limitation lies in the fact that the profiles are only based on a *snapshot* of learning *at the end of the process*. Ultimately, these behaviors are not constant across the activity, and learning is inherently characterized by episodes of both reflection-on-action and reflection-in-action (Lavoué et al., 2015) and both positive and negative emotions (Sinha, 2021). To better understand the *evolution* of these behaviors and constructs, i.e. to elaborate the *process of learning* and to further build theories of impasse-driven collaborative learning, in our future work, we aim to investigate temporal data from the same study to develop temporal understanding of the learning process. If we obtain similar findings by analyzing deeper at a temporal level using the *needed* modalities for the goal at hand, this might further strengthen the intervention framework.

Conclusion

In this paper, we investigated the link between multi-modal behaviors, performance in the task, and learning. Based on this investigation, we identified multi-modal collaborative learning profiles of dyads as they worked on an open-ended task around interactive tabletops with a robot mediator. For this, we focused in-depth on clusters generated through our previously validated approach that allows for scenarios where teams with similar learning and performance may be exhibiting different sets of multi-modal behaviors. We compared the discriminating behaviors between each pair of learner groups (*Expressive Explorers*, *Calm Tinkerers*, and *Silent Wanderers*), which helped us to identify behavioral profiles in terms of the types of exploratory interaction, the types of speech and affective behaviors that characterize gainers and *Silent Wanderers*. Since speech behaviors seem to be the most significant discriminatory factor, we also presented an interaction analysis of episodes from each group to shed light on the dialogues that took place between the team members during a phase of high overlap of speech. Finally, we identified the interplay between the different modalities and the manner in which they relate to learning and non-learning.

We must point out that our analysis in this paper is limited in certain aspects. The data driven clusters are imbalanced, meaning that with our pipeline, the non-learning cluster that emerges has fewer teams. This may be one of the reasons why the gainer profiles are clearer compared to the *Silent Wanderers* profile. Secondly, the analysis in this paper focuses on aggregate behaviors only. As future work, we are investigating how these behaviors evolve over the duration of the entire task and if similar significant differences are present among these groups at the temporal level. The eventual goal is to incorporate these insights for real-time intervention in constructivist collaborative activities.

Appendix 1

For clustering, we utilize k-means in both approaches where the value of k is selected based on entropy analysis. Approach A gives four clusters corresponding to high/low combinations of learning gains and performance metric named accordingly as *Productive Success* (high learning and performance), *Productive Failure* (high learning but low performance), *non-Productive Success* (high performance but low learning), *non-Productive Failure* (low learning and performance) abbreviated as PS, PF, non-PS, non-PF, respectively. On the other hand, approach B gives 3 behavioral clusters with the first two exhibiting high learning and the third lower learning; hence, named as *type 1 gainers*, *type 2 gainers*, and *non-gainers*, respectively.

When comparing the three behavioral clusters from *approach B*, cluster 1 and 2 both have learning gains that are significantly higher than the learning gains exhibited by the third behavioral cluster, while the average performance of all 3 behavioral clusters is very similar. When comparing the similarity between the forward and the backward clusters in terms of the teams they consist of (Appendix Fig. 15), we observe that the first two behavioral clusters have more than 70% teams from both the *Productive Failure* and *Productive Success* groups, while the third behavioral cluster mostly has teams from the *non-Productive Failure* and *non-Productive Success* groups. Concretely, this implies that learners who end up with a learning gain regardless of their performance in the task exhibit two kinds of behaviors. With the two requirements mentioned in

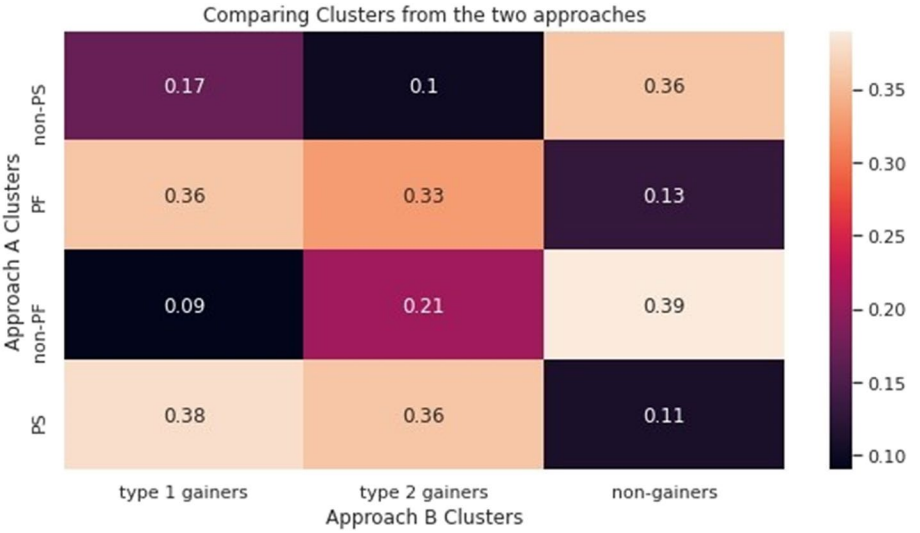


Fig. 15 Comparison between the clusters of the two approaches in terms of the teams they consist of

Table 5 Classification results

Classifier	k-fold cross-validation		Test-set	
	Accuracy	F1-score	Accuracy	Accuracy
Approach A				
SVM	0.28	0.23	0.44	0.34
RF	0.36	0.26	0.33	0.39
Approach B				
SVM	0.80	0.76	0.88	0.89
RF	0.72	0.68	0.88	0.82

"Multi-modal Learning Analytics" section being met, we can proceed with the labels surfaced from the two approaches to be used by classifiers trained on multi-modal behaviors. We made use of two commonly used classifiers, SVM and Random Forests with our dataset (Nasir, Norman, et al., 2021b). Please notice that the classifiers were trained and tested on this newer dataset version, hence providing slightly different results from those reported in Nasir, Norman, Bruno, and Dillenbourg (2020c). As can be seen in Appendix Table 5, we achieve much higher accuracy and recall on the validation and test set with labels from *approach B*; thus, lending further support to our argument that this approach is better than *approach A* in identifying the behavioral profiles of gainers and *Silent Wanderers*.

Acknowledgements This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 765955. Furthermore, we are grateful to the Swiss National Science Foundation for supporting this project through the National Centre of Competence in Research Robotics.

Code availability Not applicable.

Funding Open access funding provided by EPFL Lausanne. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 765955. Furthermore, this project is supported by the Swiss National Science Foundation through the National Centre of Competence in Research Robotics.

Data availability Not applicable.

Declarations

Conflict of interest We have no conflict of interest.

References

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human Computer Studies, 68*(4), 223–241. https://doi.org/10.1016/j.ijhcs.2009.12.003

Baltrusaitis, T., McDuff, D., Banda, N., Mahmoud, M., Kaliouby, R., Robinson, P., & Picard, R. (2011). Real-Time Inference of Mental States from Facial Expressions and Upper Body Gestures. In: (pp. 909–914). https://doi.org/10.1109/FG.2011.5771372.

Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2016). Openface: An open source facial behavior analysis toolkit, 1-10. https://doi.org/10.1109/WACV.2016.7477553.

Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences, 12*(3), 307–359.

Barron, B., Schwartz, D., Vye, N., Moore, A., Petrosino, A., Zech, L., & Bransford, J. (1998). Doing with understanding: Lessons from research on problem-and project- based learning. *Journal of the Learning Sciences, 7*(3-4), 271–311.

Bassiou, N., Tsiartas, A., Smith, J., Bratt, H., Richey, C., Shriberg, E., ... Alozie, N. (2016). Privacy-preserving speech analytics for automatic assessment of student collaboration. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *08-12-September-2016*, 888–892. https://doi.org/10.21437/Interspeech.2016-1569

Basu, S., Biswas, G., & Kinnebrew, J. S. (2017). Learner modeling for adaptive scaffolding in a computational thinking-based science learning environment. *User Modeling and User-Adapted Interaction, 27*(1), 5–53.

Benitez-Quiroz, C. F., Srinivasan, R., & Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (p. 5562-5570). https://doi.org/10.1109/CVPR.2016.600.

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. https://doi.org/10.18608/jla.2016.32.11

Campione, E., & Véronis, J. (2002). A large-scale multilingual study of pause duration. Speech prosody 2002. Proceedings of the 1st international conference on speech prosody, 199–202. Retrieved from http://www.isca-speech.org/archive/sp2002/sp02199.html

Cherubini, M., Nüssli, M.-A., & Dillenbourg, P. (2008). Deixis and gaze in collaborative work at a distance (over a shared map) a computational model to detect misunderstandings. In: *Proceedings of the 2008 symposium on eye tracking research & applications* (pp. 173–180).

Cohn, J. (2006). Foundations of human computing: Facial expression and emotion. In: *Icmi'06: 8th international conference on multimodal interfaces, conference proceeding* (pp. 233–238). https://doi.org/10.1007/978-3-540-72348-61.

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001

Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction, 22*(1-2), 9–38. https://doi.org/10.1007/s11257-011-9106-8

Dillenbourg, P., Järvelä, S., & Fischer, F. (2009). The evolution of research on computer-supported collaborative learning: From design to orchestration. In: *Technology enhanced learning* (pp. 3–19). https://doi.org/10.1007/978-1-4020-9827-7.

Dindar, M., Jarvela, S., Ahola, S., Huang, X., & Zhao, G. (2020). Leaders and followers identified by emotional mimicry during collaborative learning: A facial expression recognition study on emotional valence. *IEEE Transactions on Affective Computing, 1,* 1–1. https://doi.ieeecomputersociety.org/10.1109/TAFFC.2020.3003243.

Do-lenh, S. (2012). Supporting Reflection and Classroom Orchestration with Tangible Tabletops, *5313*, 241. https://doi.org/10.5075/epfl-thesis-5313.

Ekman, P., & Friesen, W. (1978). *Facial action coding system: Manual*. Palo Alto, Calif: Consulting Psychologists Press.

El Kaliouby, R., & Robinson, P. (2004). Real-time inference of complex mental states from facial expressions and head gestures. In: *2004 conference on computer vision and pattern recognition workshop* (p. 154–154). https://doi.org/10.1109/CVPR.2004.427.

Emara, M., Rajendran, R., Biswas, G., Okasha, M., & Elbanna, A. A. (2018). Do students' learning behaviors differ when they collaborate in open-ended learning environments? *Proceedings of the ACM on human-computer interaction*, *2*(CSCW), 1–19.

Emerson, A., Cloude, E. B., Azevedo, R., & Lester, J. (2020). Multimodal learning analytics for game-based learning. *British Journal of Educational Technology, 51*(5), 1505–1526. https://doi.org/10.1111/bjet.12992

Etkina, E., Karelina, A., Ruibal-Villasenor, M., Rosengrant, D., Jordan, R., & Hmelo-Silver, C. E. (2010). Design and reflection help students develop scientific abilities: Learning in introductory physics laboratories. *Journal of the Learning Sciences, 19*(1), 54–98. https://doi.org/10.1080/10508400903452876

Evans, A. C., Wobbrock, J. O., & Davis, K. (2016). Modeling collaboration patterns on an interactive tabletop in a classroom setting. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, 27*, 860–871. https://doi.org/10.1145/2818048.2819972

Fors, K. L. (2015). Production and perception of pauses in speech (Doctoral dissertation, University of Gothenburg). Retrieved from https://gupea.ub.gu.se/bitstream/2077/39346/1/gupea_2077_39346_-1.pdf

Giannakos, M. N., Sharma, K., Pappas, I. O., Kostakos, V., & Velloso, E. (2019). Multimodal data as a means to understand the learning experience. *International Journal of Information Management, 48*(March), 108–119. https://doi.org/10.1016/j.ijinfomgt.2019.02.003

Hayashi, Y. (2019). Detecting collaborative learning through emotions: An investigation using facial expression recognition. In: *International conference on intelligent tutoring systems* (pp. 89–98).

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics, 38*(4), 555–568. https://doi.org/10.1016/j.wocn.2010.08.002

Herman, A. M., Critchley, H. D., & Duka, T. (2018). The role of emotions and physiological arousal in modulating impulsive behaviour. *Biological Psychology, 133*, 30–43. Retrieved from https://www.sciencedirect.com/science/article/pii/S0301051118300644. https://doi.org/10.1016/j.biopsycho.2018.01.014

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review, 16*(3), 235–266.

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42*(2), 99–107. https://doi.org/10.1080/00461520701263368

Hone, K. (2006). Empathic agents to reduce user frustration: The effects of varying agent characteristics. *Interacting with Computers, 18*(2), 227–245. https://doi.org/10.1016/j.intcom.2005.05.003

Huang, K., Bryant, T., & Schneider, B. (2019). Identifying collaborative learning states using unsupervised machine learning on eye-tracking, physiological and motion sensor data. *EDM 2019 - Proceedings of the 12th International Conference on Educational Data Mining*(Edm), 318–323.

Järvelä, S., & Hadwin, A. F. (2013). New frontiers: Regulating learning in CSCL. *Educational Psychologist, 48*(1), 25–39.

Järvelä, S., Kirschner, P. A., Hadwin, A., Järvenoja, H., Malmberg, J., Miller, M., & Laru, J. (2016). Socially shared regulation of learning in CSCL: Understanding and prompting individual- and group-level shared regulatory activities. *International Journal of Computer-Supported Collaborative Learning, 11*, 263–280. https://doi.org/10.1007/s11412-016-9238-2

Järvelä, S., Gašević, D., Seppänen, T., Pechenizkiy, M., & Kirschner, P. A. (2020). Bridging learning sciences, machine learning and affective computing for understanding cognition and affect in collaborative learning. *British Journal of Educational Technology, 51*(6), 2391–2406.

Jermann, P., & Nüssli, M.-A. (2012). Effects of sharing text selections on gaze cross-recurrence and interaction quality in a pair programming task. In: *Proceedings of the ACM 2012 conference on computer supported cooperative work* (pp. 1125–1134).

Jermann, P., Mullins, D., Nüssli, M.-A., & Dillenbourg, P. (2011). Collaborative gaze footprints: Correlates of interaction quality. In: *Connecting computer-supported collaborative learning to policy and practice: Cscl2011 conference proceedings.* (pp. 184–191).

Kapur, M. (2008). Productive failure. *Cognition and Instruction, 26*(3), 379–424. https://doi.org/10.1080/07370000802212669

Kapur, M. (2011). Temporality matters: Advancing a method for analyzing problem-solving processes in a computer-supported collaborative environment. *International Journal of Computer-Supported Collaborative Learning, 6*(1), 39–56.

Kauschke, C., Bahn, D., Vesker, M., & Schwarzer, G. (2019). The role of emotional valence for the processing of facial and verbal stimuli—Positivity or negativity Bias? *Frontiers in Psychology, 10*, 1654. https://doi.org/10.3389/fpsyg.2019.01654

Kim, J., Truong, K. P., Charisi, V., Zaga, C., Lohse, M., Heylen, D., & Evers, V. (2015). Vocal turn-taking patterns in groups of children performing collaborative tasks: An exploratory study. In: *Interspeech 2015* (pp. 1645–1649).

Kinnebrew, J. S., Loretz, K. M., & Biswas, G. (2013). A contextualized, differential sequence mining method to derive students' learning behavior patterns. *Journal of Educational Data Mining, 5*(1), 190–219.

Kirschner, P., Sweller, J., & Clark, R. E. (2006). Why unguided learning does not work: An analysis of the failure of discovery learning, problem-based learning, experiential learning and inquiry-based learning. *Educational Psychologist, 41*(2), 75–86.

Kirschner, F., Paas, F., & Kirschner, P. A. (2011). Task complexity as a driver for collaborative learning efficiency: The collective working-memory effect. *Applied Cognitive Psychology, 25*(4), 615–624. https://doi.org/10.1002/acp.1730

Klein, J., Moon, Y., & Picard, R. (2002). This computer responds to user frustration: Theory, design, and results. *Interacting with Computers, 14*(2), 119–140. https://doi.org/10.1016/S0953-5438(01)00053-4

Land, S. M., Hannafin, M. J., & Oliver, K. (2000). Student-centered learning environments. *Theoretical Foundations of Learning Environments*, 2nd Ed., Chapter 1, pp. 1–23.

Lavoué, É., Molinari, G., Prié, Y., & Khezami, S. (2015). Reflection-in-action markers for reflection-on-action in computer-supported collaborative learning settings. *Computers & Education, 88*, 129–142.

Liu, R., Stamper, J. C., & Davenport, J. (2018). A novel method for the in-depth multimodal analysis of student learning trajectories in intelligent tutoring systems. *Journal of Learning Analytics, 5*(1), 41–54. https://doi.org/10.18608/jla.2018.51.4

Lodge, J. M., Kennedy, G., Lockyer, L., Arguel, A., & Pachman, M. (2018). Understanding difficulties and resulting confusion in learning: An integrative review. *Frontiers in Education, 3*, 1–10. https://doi.org/10.3389/feduc.2018.00049

Loibl, K., & Rummel, N. (2014). The impact of guidance during problem-solving prior to instruction on students' inventions and learning outcomes. *Instructional Science, 42*(3), 305–326.

Loibl, K., Roll, I., & Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. *Educational Psychology Review, 29*(4), 693–715. https://doi.org/10.1007/s10648-016-9379-x

Lou, Y., Abrami, P. C., & d'Apollonia, S. (2001). Small group and individual learning with technology: A meta-analysis. *Review of Educational Research, 71*(3), 449–521.

Malmberg, J., Haataja, E., Seppänen, T., & Järvelä, S. (2019a). Are we together or not? The temporal interplay of monitoring, physiological arousal and physiological synchrony during a collaborative exam. *International Journal of Computer-Supported Collaborative Learning, 14*(4), 467–490.

Malmberg, J., Järvelä, S., Holappa, J., Haataja, E., Huang, X., & Siipo, A. (2019b). Going beyond what is visible: What multichannel data can reveal about interaction in the context of collaborative learning? *Computers in Human Behavior, 96*, 235–245. https://doi.org/10.1016/j.chb.2018.06.030

Maroni, B., Gnisci, A., & Pontecorvo, C. (2008). Turn-taking in classroom interactions: Overlapping, interruptions and pauses in primary school. *European Journal of Psychology of Education, 23*(1), 59–76. https://doi.org/10.1007/BF03173140

Martinez, R., Wallace, J. R., Kay, J., & Yacef, K. (2011). Modelling and identifying collaborative situations in a collocated multi-display groupware setting. In: *International conference on artificial intelligence in education* (pp. 196–204).

Martinez-Maldonado, R., Dimitriadis, Y., Martinez-Monés, A., Kay, J., & Yacef, K. (2013). Capturing and analyzing verbal and physical collaborative learning interactions at an enriched interactive tabletop. *International Journal of Computer-Supported Collaborative Learning, 8*(4), 455–485.

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.

Mentis, H. M., et al. (2007). Memory of frustrating experiences. In: D. Nahl, & D. Bilal (Eds.), *Information and emotion: The emergent affective paradigm in information behavior research and theory* (pp. 197–210). Information Today, Inc.

Nasir, J., Norman, U., Johal, W., Olsen, J., Shahmoradi, S., & Dillenbourg, P. (2019). Robot analytics: What Do human-robot interaction traces tell us about learning? 2019. 28th *IEEE international conference on robot and human interactive communication* (Roman), 1–7.

Nasir, J., Bruno, B., & Dillenbourg, P. (2020a). Is there 'one way' of learning? A data-driven approach. In: *Companion publication of the 2020 international conference on multimodal interaction* (p. 388–391). New York: Association for Computing Machinery. https://doi.org/10.1145/3395035.3425200.

Nasir, J., Norman, U., Bruno, B., Chetouani, M., & Dillenbourg, P. (2020b). PE-HRI: A multimodal dataset for the study of productive engagement ina robot mediated collaborative educational setting. *Zenodo*. https://doi.org/10.5281/zenodo.4288833

Nasir, J., Norman, U., Bruno, B., & Dillenbourg, P. (2020c). When positive perception of the robot has no effect on learning. In: *IEEE International Conference on Robot and Human Interactive Communication (Roman).*

Nasir, J., Bruno, B., Chetouani, M., & Dillenbourg, P. (2021a). What if social robots look for productive engagement*? International Journal of Social Robotics*. https://doi.org/10.1007/s12369-021-00766-w

Nasir, J., Norman, U., Bruno, B., Chetouani, M., & Dillenbourg, P. (2021b). PE-HRI: A multimodal dataset for the study of productive engagement in arobot mediated collaborative educational setting. *Zenodo*. https://doi.org/10.5281/zenodo.4633092

Norman, U., Dinkar, T., Nasir, J., Bruno, B., Clavel, C., & Dillenbourg, P. (2021). Justhink dialogue and actions corpus. *Zenodo*. https://doi.org/10.5281/zenodo.462710

Olsen, J. K., Sharma, K., Rummel, N., & Aleven, V. (2020). Temporal analysis of multi-modal data to predict collaborative learning outcomes. *British Journal of Educational Technology, 51*(5), 1527–1547. https://doi.org/10.1111/bjet.12982

Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaïane, O. R. (2008). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering, 21*(6), 759–772.

Pijeira-díaz, H. J., Drachsler, H., Järvelä, S., & Kirschner, P. A. (2019). Sympathetic arousal commonalities and arousal contagion during collaborative learning : How attuned are triad members? *Computers in Human Behavior, 92*, 188–197. https://doi.org/10.1016/j.chb.2018.11.008

Popov, V., van Leeuwen, A., & Buis, S. (2017). Are you with me or not? Temporal synchronicity and transactivity during cscl. *Journal of Computer Assisted Learning, 33*(5), 424–442.

Praharaj, S., Scheffel, M., Drachsler, H., & Specht, M. (2021). Literature review on co-located collaboration modeling using multimodal learning analytics — Can we go the whole nine yards? *IEEE Transactions on Learning Technologies, 14*(3), 367–385. https://doi.org/10.1109/TLT.2021.3097766

Reilly, J. M., & Schneider, B. (2019). Predicting the quality of collaborative problem solving through linguistic analysis of discourse. *EDM 2019 - Proceedings of the 12th International Conference on Educational Data Mining*(Edm), 149–157.

Rodríguez, F. J., & Boyer, K. E. (2015). Discovering individual and collaborative problem-solving modes with hidden Markov models. In: *Artificial intelligence in education: Proceedings of the world conference on AI in education 2015* (pp. 408–418). https://doi.org/10.1007/978-3-319-19773-9.

Roschelle, J. (1992). Learning by collaborating: Convergent conceptual change. *Journal of the Learning Sciences, 2*(3), 235–276. https://doi.org/10.1207/s15327809jls0203-1

Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In: *Computer supported collaborative learning* (pp. 69–97).

Rummel, N., & Spada, H. (2005). Learning to collaborate: An instructional approach to promoting collaborative problem solving in computer-mediated settings. *The Journal of the Learning Sciences, 14*(2), 201–241.

Russell, J. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*, 145–172. https://doi.org/10.1037//0033-295X.110.1.145

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.

Schneider, B., & Pea, R. (2015). Does seeing one another's gaze affect group dialogue? A computational approach. *Journal of Learning Analytics, 2*(2), 107–133.

Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., & Pea, R. (2016). Using mobile eye-trackers to unpack the perceptual benefits of a tangible user interface for collaborative learning. *ACM Transactions on Computer-Human Interaction, 23*(6), Article No.: 39. https://doi.org/10.1145/3012009

Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., & Pea, R. (2018). Leveraging mobile eye-trackers to capture joint visual attention in co-located collaborative learning groups. *International Journal of Computer-Supported Collaborative Learning, 13*(3), 241–261.

Schneider, B., Dich, Y., & Radu, I. (2020). Unpacking the relationship between existing and new measures of physiological synchrony and collaborative learning: A mixed methods study. *International Journal of Computer-Supported Collaborative Learning, 15*(1), 89–113.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16*(4), 475–5223.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction, 22*(2), 129–184.

Sharma, K., Caballero, D., Verma, H., Jermann, P., & Dillenbourg, P. (2015). Looking at versus looking through: A dual eye-tracking study in mooc context. In: *Proceedings of 11th international conference of computer supported collaborative learning* (Vol. 1, pp. 260–267).

Sharma, K., Papamitsiou, Z., Olsen, J. K., & Giannakos, M. (2020). Predicting learners' effortful behavior in adaptive assessment using multimodal data. *ACM International Conference Proceeding Series*, 480–489. https://doi.org/10.1145/3375462.3375498

Sharma, K., Olsen, J. K., Aleven, V., & Rummel, N. (2021). Measuring causality between collaborative and individual gaze metrics for collaborative problem-solving with intelligent tutoring systems. *Journal of Computer Assisted Learning, 37*(1), 51–68.

Sinha, T. (2021). Enriching problem-solving followed by instruction with explanatory accounts of emotions. *Journal of the Learning Sciences*, 1–48.

Spikol, D., Ruffaldi, E., & Cukurova, M. (2017). Using multimodal learning analytics to identify aspects of collaboration in project-based learning. *Computer-Supported Collaborative Learning Conference, CSCL, 1*, 263–270. https://doi.org/10.22318/cscl2017.37

Spikol, D., Ruffaldi, E., Dabisias, G., & Cukurova, M. (2018). Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *Journal of Computer Assisted Learning, 34*(4), 366–377.

Stahl, G., Law, N., & Hesse, F. (2013). Reigniting CSCL flash themes. *International Journal of Computer-Supported Collaborative Learning, 8*(4), 369–374. https://doi.org/10.1007/s11412-013-9185-0

Teasley, S. D. (1997). Talking about reasoning: How important is the peer in peer collaboration? In: *Discourse, tools and reasoning* (pp. 361–384). Springer.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction, 21*(3), 209–249.

Veenman, M. V. J. (2013). Assessing metacognitive skills in computerized learning environments. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 157–168). Springer. https://doi.org/10.1007/978-1-4419-5546-3-11

Viswanathan, S. A., & VanLehn, K. (2017). Using the tablet gestures and speech of pairs of students to classify their collaboration. *IEEE Transactions on Learning Technologies, 11*(2), 230–242.

Vrzakova, H., Amon, M. J., Stewart, A., Duran, N. D., & D'Mello, S. K. (2020). Focused or stuck together: Multimodal patterns reveal triads' performance in collaborative problem solving. *ACM International Conference Proceeding Series: Learning Analytics and Knowledge, 2020*, 295–304. https://doi.org/10.1145/3375462.3375467

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education, 46*(1), 71–95.

Worsley, M., & Blikstein, P. (2011). What's an expert? Using learning analytics to identify emergent markers of expertise through automated speech, sentiment and sketch analysis. In: *Edm* (pp. 235–240).

Worsley, M., & Blikstein, P. (2018). A multimodal analysis of making. *International Journal of Artificial Intelligence in Education, 28*, 385–419.

Authors and Affiliations

Jauwairia Nasir[1] · Aditi Kothiyal[1] · Barbara Bruno[1] · Pierre Dillenbourg[1]

Aditi Kothiyal
aditi.kothiyal@epfl.ch

Barbara Bruno
barbara.bruno@epfl.ch

Pierre Dillenbourg
pierre.dillenbourg@epfl.ch

[1] Computer-Human Interaction in Learning and Instruction (CHILI) Lab, Swiss Federal Institute of Technology in Lausanne (EPFL), Lausanne, Switzerland