Estimation of ICAP States Based on Interaction Data During Collaborative Learning

Yoshimasa Ohmoto
Shizuoka University
Hamamatsu, Japan
ohmoto-y@inf.shizuoka.ac.jp

Shigen Shimojo
Ristumeikan University
Ibaraki, Japan
sshimojo@fc.ritsumei.ac.jp

Junya Morita
Shizuoka University
Hamamatsu, Japan
j-morita@inf.shizuoka.ac.jp

Yugo Hayashi
Ristumeikan University
Ibaraki, Japan
yhayashi@fc.ritsumei.ac.jp

The primary goal of this study is to investigate a method for estimating the state of learners in the near future using nonverbal information used in multimodal interaction as cues to provide adaptive support in collaborative learning. We used interactive-constructive-active-passive (ICAP) theory to classify learners' states in collaborative learning. We attempted to determine whether a learner's ICAP state was passive based on multimodal data obtained during a collaborative concept-map task. We conducted an experiment on collaborative learning among learners and acquired data on conversational type, the results of learning performance (pre- and post-tests), utterances, facial expressions, gaze, and voice during the experiment. We conducted two analyses. One was sequential pattern mining, to obtain clues for predicting the participants' state after 5 seconds. The other was a support vector machine to try to classify the participants' state based on the obtained clues. We found several candidates that could be used for learner-state estimation in the near future. The learner-state estimation using multimodal information yielded higher than 70% accuracy. In contrast, there were differences in the ease of estimating each pair's learning state. It appears that capturing the characteristics of interactions in collaborative learning for each pair is necessary for a more accurate estimation of the learners' state.

Keywords: multimodal information, learner state estimation, human behavior analysis, collaborative learning, concept map, ICAP theory, Computer-Supported Collaborative Learning (CSCL)

1. INTRODUCTION

Constructive interactions in collaborative learning are known to be effective in promoting one's understanding by incorporating different perspectives (Chi et al., 1994; Okada and Simon, 1997; Shirouzu et al., 2002). Collaborative learning requires the the presence of appropriate attitudes on the part of both instructors and participants. However, the number of instructors is limited, and it is difficult for participants with little experience to have an appropriate attitude. Therefore, computer support has been proposed as a solution to these issues. For example, computer-supported collaborative learning (CSCL) is a form of learning support that has been

researched for a long time (Bodemer, 2011; Rummel et al., 2008; Dillenbourg and Fischer, 2007). These studies have provided a number of insights into collaborative learning, including the development of methods to support communication between participating learners and foster deeper understanding. In addition, intelligent tutoring systems (ITSs) using pedagogical conversational agents (PCAs) have also been proposed as a form of computer support (Carbonell, 1970; Hayashi, 2019a). ITSs using PCAs can conduct metacognitive facilitation and provide adaptive feedback, which enhances the collaborative learning process in CSCL environments (Koedinger et al., 1997). Recent studies have attempted to use techniques such as rule-based reasoning, data-mining-based methods, and Bayesian networks to provide better support within these systems (Mousavinasab et al., 2021). These advanced techniques can be integrated into CSCL platforms to enrich the learning experience by offering tailored guidance and feedback to participants.

The ultimate goal of this study is to develop a collaborative learning environment that is beneficial to learners by involving PCAs as both instructors and experienced participants in collaborative learning. Therefore, it is necessary to construct a dynamic interaction model between a learner and a PCA by clarifying predictors that are clues for supporting collaborative learning. In this study, we examine the usefulness of multimodal information as clues for providing adaptive support based on learners' states in collaborative learning. One of the predictors is the internal state of the learner in collaborative learning. This includes the learner's basic knowledge, emotions, planning of task, and comprehension of the task. All of these factors are important, but in this study, we focused on the mental attitude of learners toward the interaction in the task as "learner's state" as a preliminary step to implement in ITS and PCA. The estimation of such mental attitude provides a clue to proactively predict learners' behavior and reactions in the near future. To investigate this, we try to obtain data on collaborative learning interactions between humans and attempt to predict the "learner's state" in the near future.

1.1. Computer Support in Collaborative Learning

Appropriate constructive interactions between participants are essential for the success of collaborative learning (Miyake, 1986; Shirouzu et al., 2002; Chi, 2009). Studies in cognitive science have explored the process of successful collaboration between participants with different perspectives (Hayashi et al., 2007; Hayashi, 2018). Meanwhile, in environments where communication channels are constrained, including computer-mediated interactions such as text chat, and it is difficult to estimate the state between learners based on multimodal information, collaborative problem-solving based on different perspectives has been found to be more likely to fail to build a common ground (Hayashi and Miwa, 2011).

Multimodal interactions refer to interactions in which the participants use multiple types of verbal and nonverbal information that learners consciously or unconsciously use in their interactions with others. Our natural communication is multimodal, comprising both speech and such bodily behaviors as facial expressions, eye gaze, and co-speech gestures. Despite the limitations of the CSCL environment to support multimodal interactions, users use the interaction abilities they leverage in their daily lives to engage with other users. The multimodal behavior can provide information about the shared and temporal aspects of collaborative learning and help design models and tools for more effective collaborative learning (Järvelä et al., 2022).

In collaborative learning, learners form a common ground by estimating the elements they are paying attention to and the content of tasks they understand. Multimodal interaction is expected to play an important role in the formation of such a common ground. The importance of

this has also been considered in CSCL, and attempts have been made to use social-recognition tools to reduce the constraints on communication channels, promote joint attention, and help establish a common ground for interaction (Dillenbourg and Fischer, 2007; Bodemer, 2011; Belenky et al., 2014). Therefore, we focused on estimating the learner's states based on information used in multimodal interaction between learners as cues.

Among the challenges in CSCL, one is to determine whether a learner needs to be supported in the process of collaborative learning and to dynamically provide interventions. Recent research has suggested that interactions based on learners' understanding and attitudes toward their tasks are important (e.g., Hayashi 2019b; Hayashi 2019a). In this research field, a PCA is sometimes used to monitor learners' behavior and intervene when they need support or reflection from a third party. Therefore, how agents present information via prompts and interventions in collaborative learning to facilitate learners' conversations, such as dialog with explanations, is being examined. For example, Rummel and Spada (2005) reported that a collaboration script as an instructional method to promote graduate students' ability to collaborate well showed positive effects on three levels: (a) the collaborative process, (b) the outcomes they obtained jointly, and (c) their individual knowledge about characteristic features of a good collaboration assessed in a posttest. ITSs implemented with cognitive models have also been investigated, and methods for facilitating learners' cognitive processes, such as self-reflection, during interactive knowledge teaching with learners have been proposed (Graesser and McNamara, 2010). Tutorial dialog systems using conversational agents have also been proposed and shown to be effective in a variety of fields (O'Neill et al., 2003; Jordan et al., 2006; VanLehn et al., 2007). Such systems can help facilitate more active learner's states because they can elicit explanations on a step-by-step basis so that each step of reasoning contributes to the resulting explanation.

However, these methods do not always have a positive effect. Rummel et al. (2009) examined the effects of collaboration scripts in detail and found mixed results. Various adaptive supports have been proposed for these problems (Gweon et al., 2006; Walker et al., 2008). A support method that provides such adaptive feedback is a cognitive tutor (Anderson et al., 1995), a type of ITS. Aleven and Koedinger (2002) proposed cognitive tutors that promoted metacognitive activities through self-explanatory activities as one of the learning aids. In this method, a premade model was used to detect a learner's state and provided relevant and appropriate feedback. Meanwhile, how to provide adaptive support is still an issue to be investigated. An important issue in addressing this is the real-time estimation of the learner's state during collaborative learning.

1.2. Learners' Interaction States during Collaborative Learing

The interactive-constructive-active-passive (ICAP) theory has been proposed as one of the frameworks for classifying learners' interaction states in collaborative learning (Chi, 2009; Chi and Wylie, 2014). The ICAP theory describes learners' activity states during collaborative learning based on the following four states. 1) Passive state: passively learning from a teacher and learning materials. 2) Active state: actively and voluntarily tackling problems with explicit action. This activity often involves some physical action or manipulation by the learner and/or intellectual activities such as examining a problem from different perspectives. 3) Constructive state: a state of activity in which some additional element is added to the learning material presented to the learner to "generate" the knowledge or representation for it. This activity also includes activities in which the learner externalizes knowledge by verbalizing the content of the knowledge

in his/her head through self-explanation activities. 4) Interactive state: the reconstruction of knowledge through criticism and refutation of interaction partners' ideas. The content presented in the dialogue must contribute to the generation of the knowledge that is ultimately generated. This activity involves a dialogue in which different points of view are expressed, rationales are given, explanations are sought, and questions are asked about the other person's knowledge and viewpoints.

It has been shown that as a learner's state transitions to the interactive state, their performance in collaborative learning heightens (I > C > A > P) (Chi, 2009). For example, Wiggins et al. (2017) quasi-empirically examined whether the ICAP framework could predict learning performance in science, technology, entrepreneurship, and mathematics (STEM) classes. Subsequently, the instructor indicated that learners made more efforts to support interactive activities because of their need to ensure sufficient time.

If a learner simply responds to another's question, they are in the passive state, even if the learner is efficiently searching for a solution. This state may be beneficial to the questioner; however, it is not beneficial to the respondent. When working on a task with a PCA, the interactive state should be targeted based on the ICAP theory. However, in current human-agent interactions, learners often remain in the passive state (Raux et al., 2005; Misu et al., 2011). This is a problem not only for agents in collaborative learning but also for interactive agents in general. However, to induce a learner to a state other than the passive, it is necessary to elicit the learner's spontaneity; thus, it is not sufficient to follow a specific procedure or provide particular information. It is necessary to consider ways to elicit the learner's spontaneity by adaptively providing feedback according to the learner's state. This research is premised on our conjecture of a significant difference between the passive and active states in the ICAP theory in terms of a learner's spontaneity, and we considered whether we could discriminate the passive from the other states.

1.3. Nonverbal Cues for Estimating the Learner's States

To estimate a learner's state based on information that can be obtained mechanically, several types of information can be considered as cues. In previous research on CSCL, the effectiveness of collaborative learning was found to largely depend on the richness and intensity of the interactions in which group members engaged during the collaboration process (Dillenbourg et al., 1996). As one of the indicators of the richness of interaction, attempts have been made to estimate a learner's state by analyzing text chats (Wang et al., 2015). The contents of learners' conversations are the most promising cue. However, text chats take time to create inputs and may not be suitable for the quick sharing of learners' thoughts.

In CSCL environments, the history of the use of the tools available in that environment also provides cues for estimating a learner's state. For example, studies using a shared concept map in which learners could observe the knowledge state of other learners participating in collaborative learning demonstrated the ability of learners to jointly acquire new knowledge by analyzing changes in the created concept map (Engelmann and Hesse, 2010; Sangin et al., 2011). A concept map is useful for understanding the state of knowledge held by learners; however, it may also play an important role in promoting collaborative learning as new perspectives are gained by referring to the perspectives of others. Not only the concept map but also the tools in the CSCL environments often play various roles. Therefore, to utilize operation logs to estimate a learner's state, it is necessary to have a deep understanding of the role the tools play in collaborative

learning.

In this study, we focused on multimodal interaction, which is thought to be useful for human state estimation. Blikstein and Worsley (2016) reviewed the importance of learning analytics using multimodal information. The importance of nonverbal information has also been pointed out in the labeling of each state in ICAP. For example, Chi and Wylie (2014) and Chi et al. (2018) argued that it is necessary to estimate a learner's state by taking into account nonverbal information that occurs during the interaction, such as overt motor action and physical manipulation, pointing or gesturing at parts of what learners are reading or solving, or pausing and rewinding parts of a videotape for review. Several studies have proposed assessment methods based on the ICAP framework to analyze student learning activities (Goldberg et al., 2021; Vosniadou et al., 2023; Hsiao et al., 2022). These studies also used not only verbal but also multimodal cues to label students' actions, suggesting the importance of multimodal interaction in learning activities applied ICAP theory.

Stewart et al. (2021) examined a multimodal model in collaborative problem-solving, where the multimodal information used was language, acoustic-prosodic features, facial features, body movement, and task information. Using a random forest and a neural network which combined verbal and nonverbal models, they were successful in classifying the three core collaborative problem-solving facets: construction of shared knowledge, negotiation/coordination, and maintaining team function. This model is useful, but it makes use of utterance content, which is problematic due to errors in automatic transcription and processing time. Kasparova et al. (2021) attempted to classify students' involvement in collaborative problem-solving from video data using LSTM into three categories: active, semi-active, and passive. There is no semi-active category in ICAP, and the "semi-active" cases would be classified as either Active or Passive in the ICAP classification, depending on how the learner is performing the intellectual activity. Overall, the classification accuracy was not very high, but in some cases, it was partially successful. Ma et al. (2022) attempted to detect impasses in collaborative problem-solving using language, phonological information, and facial expressions. The final classification result was more than 80% using multimodal information. The results are important, but the problem remains that the unit of analysis was an utterance, so it does not assume that the learner's state changes in the silent, and that it takes time to process to analyze the contents of utterances.

We focused on the following multimodal information. Since we aim to contribute to the realization of ITS and PCA that respond in real time, we did not analyze the contents of utterances that require processing time in the explanatory variables. We also aim to estimate the learner's states in the near future from information used in multimodal interaction.

Gaze behavior has been studied in the context of social relationships and interpersonal interactions. Previous studies (Schneider and Pea, 2013; Schneider and Pea, 2014) have shown that visual representations of a partner's gaze during a computer-based remote-learning task can promote social cooperation and learning. Hayashi (2020) showed that sharing participants' perspectives in collaborative learning positively impacted the process of collaborative learning through an experiment in which participants' perspectives were displayed on a monitor during collaborative learning. These results suggest that eye movement can be used as a cue to estimate a learner's state and adaptively respond to the learner.

Facial expressions have been studied for many years, and it is established that they can be used to estimate human emotions (Ekman et al., 1980). The facial-action coding system is a comprehensive system that uses facial excitation muscles to identify facial movements (Ekman and Friesen, 1976). This system can encode facial expressions on the human face into a

machine-discriminable form by combining the presence or absence of movements of various facial muscles called action units (AUs). Previous research has reported that facial expressions may help estimate emotions related to mutual understanding during collaborative learning (Hayashi, 2019a). Cai et al. (2020) analyzed facial expressions in ICAP states by observing learners' AU movements, though they did not investigate the passive state because the movements involved no utterance and were not clear. The results suggested that facial expressions could be a cue for estimating a learner's ICAP state.

Speech (prosodic) information is another feature that is often used in estimating human emotions and intentions. Moreover, many studies have been conducted on mechanical classification because speech information is easy to mechanically acquire and analyze in a laboratory environment (e.g., Cho and Kato 2011; Sim et al. 2002; Iliou and Anagnostopoulos 2010). These studies often use fundamental frequency (F0), energy (Intensity), and formants (F1, F2, ...) as features. Prosodic information has also been said to contain cues for estimating assertiveness and confidence (Scherer et al., 1973). These features were thought to be cues for estimating a learner's state.

In this study, we attempted to estimate a learner's state using the above characteristics. While it is important to know a learner's current state, we considered it also important to estimate the state to which the learner's state was changing in the near future. Therefore, we also analyzed time-series patterns of features.

1.4. Goal of This Study

Our future goal is to develop an ITS using a PCA that provides appropriate adaptive feedback based on a learner's state predicted based on multimodal data. The difficulties that must be overcome to achieve this goal are numerous. Among them, this study focused on examining cues to determine whether advice or intervention is needed as a preliminary study in providing adaptive feedback and in estimating the ICAP framework-based learner's states.

Previous studies to estimate learner states based on the ICAP framework have often estimated learner's states at a given point in time by referring to the content of utterances. For example, Wang et al. (2015) estimated the state of the ICAP from text of posts in MOOC discussion forums and examined the relationship between the quantity and quality of student participation and learning effectiveness. Although several studies have used multimodal information to assess learner status and learning based on the ICAP framework, the content of speech often plays an important role (Goldberg et al., 2021; Vosniadou et al., 2023; Hsiao et al., 2022). Our study is novel in two points: 1) we attempted to estimate the learner's state in the near future rather than estimating the learner's state when the learner was active, and 2) we attempted to classify passive states and non-passive states (active, constructive, or interactive [ACI states]) without depending on the contents of the utterance.

It is important to interact with other learners in collaborative learning. In general, interaction involves not only reactive responses based on information available at a given point in time but also actions to change a future state into a specific state based on one's objectives. Directing learners into a desired state involves not only providing feedback based on the learner's state at a given point in time but also proactively predicting the learner's behavior and determining whether the learner's state in the future should be changed into a desired state by intervention.

When learner's state estimation methods that rely on the contents of utterances are applied to near-future learner state estimation, there are several situations in which false positives are

expected to increase. For example, it is common for learners to be engaged in various intellectual activities even when they are not talking, and in these situations, the learner's state can be considered an Active state. Also, while they are listening in a conversation, they do not speak but are considered to be in a mixture of Active and Passive states. Furthermore, learners often do not speak just before the learner's state changes to ACI in the near future. Being considered passive by ITS or PCA and receiving an intervention despite being in an active state may have a negative impact on the learner's motivation and the learner's feeling of being denied the activity they were doing at the time. To avoid this, it is important to determine whether an intervention should be performed immediately before starting the intervention, without depending on the content of the utterance to determine whether the learner's state is passive or ACI.

First, we tried to detect whether a learner's ICAP state in the near future was passive from multimodal data obtained during collaborative learning. We reported on a re-analysis of previous experimental data in ITS2021 (Ohmoto et al., 2021). In the current study, we conducted an additional experiment on collaborative learning to enlarge the dataset and acquired data on conversational type, the results of learning performance (pre- and post-tests), facial expressions, gaze, and speech during the experiment. We labeled the ICAP states from the experiment video and searched for cues to classify the passive state of the ICAP in the near future based on the acquired data.

2. METHOD

2.1. PARTICIPANTS

The study was conducted after an institutional ethical review and approval by the ethics review committee of the (co-)authors' university. There were 24 study participants (females: 15, males: 9; average age: 19.13, *SD*=0.60), and all were Japanese students majoring in psychology; they participated through a participant pool in exchange for course credit. Only freshmen and sophomores majoring in psychology participated and were randomly grouped into dyads. The experimenter confirmed that students within a dyad had not previously interacted with one another. They tackled the task independently.

2.2. EXPERIMENT SYSTEM

The experimental system used is shown in Figure 1. Two personal computers and two monitors were used by the participants; two Sony HDRCX680 video recorders filmed their conversations and facial expressions, and two Tobii eye trackers (X2-30; https://www.tobiipro.com/) recorded their eye gaze. The video recorder and eye tracker both acquired data at 30 Hz. The audio was recorded at 44,100 Hz. A trigger signal was input to the video data and eye tracker, and time synchronization was performed based on this signal. All data, including the eye tracker, was measured without contact. Cmap tools (https://cmap.ihmc.us/) were used in the experimental task because each of the participants made an individual and a collaborative concept map. Participants were shown a video of using Cmap tools to create a concept map. Participants were able to ask questions until they understood how to use Cmap tools and what they had to do to complete the task. A monitor and video recorder were placed in front of the participants in pairs. They sat across from each other with a partition placed between them, and thus they could not see each other.

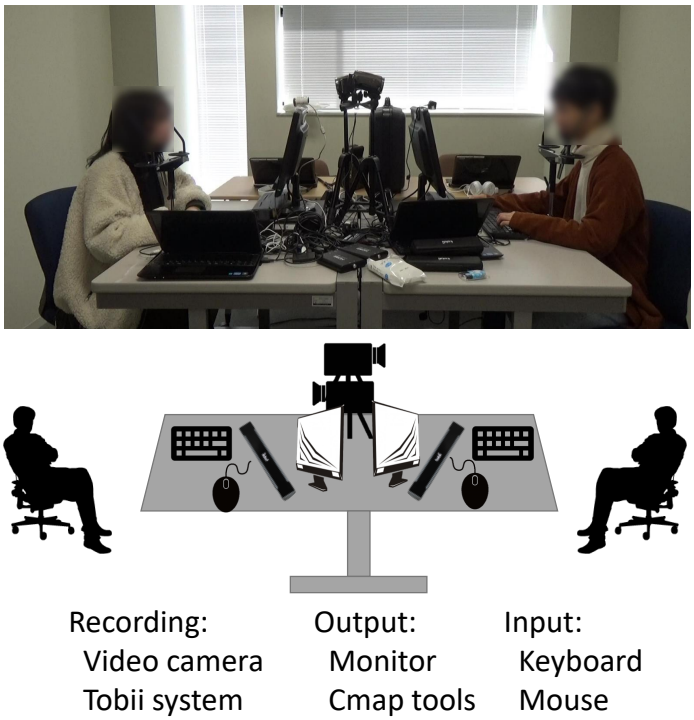

Figure 1: Experiment setup. The learners sit at the same table but cannot see each other.

An example screen displayed on each participant's monitor during the collaborative work is shown in Figure 2. The screen's right half showed the concept maps created by each participant before the collaborative work; these concept maps could not be edited during the collaborative work. A concept map to be collaboratively created was displayed on the screen's left half. This concept map could be edited by both participants during the collaborative phase.

2.3. TASK

The purpose of the experiment was to observe changes in ICAP state through human-human interaction in the absence of assistance from a third party (e.g., ITS or PCA) and to obtain clues to predict them. Therefore, participants created concept maps using Cmap, but the system they used did not implement intelligent support mechanisms.

The entire experimental procedure is as shown in Figure 3, which includes two experimental tasks in consecutive phases. The first of these was the **individual task**. The participants were instructed to individually read a text provided on a key psychological term (attribution theory). This text was prepared based on Weiner (1985) and Nasu (1988). In attribution theory, there are three axes: internal/external, stable/unstable, and controllable/uncontrollable. The overall objective of the experimental task was to explain the causal attribution of success and failure based on these axes. After confirming that they had read the text, they were instructed to read about a related episode. This episode is used in Weinberger and Fischer (2006), and we translated it into Japanese. The episode was set as follows: "You are a high-school education intern participating in a school counseling session with Michael Peter, a first-year high-school student. Please consider the causal attribution of Michael Peter's 'back-to-school worries'. The charac-

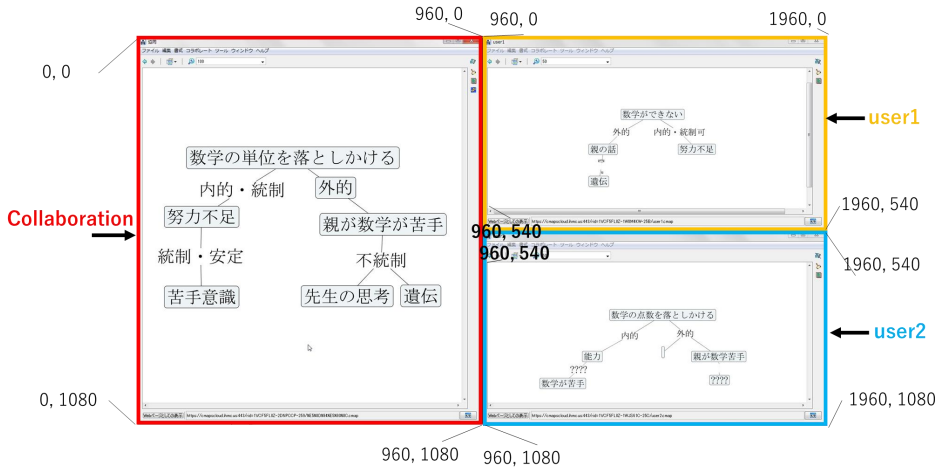

Figure 2: This is an sample screen which displayed on each participant's monitor during the collaborative work.

ters (his father, mother, and math teacher) participated in school counseling with Michael Peter, who discussed his worries about the new semester."

After reading the episode, the two participants needed to explain why the student (Michael Peter) was worried about the new semester based on the text about causal attribution and were instructed to separately create a concept map that explained causal attribution using Cmap. In this Individual phase, participants were in the same room but did not interact with each other.

The second experimental task was the **collaboration task**. Both participants were instructed to work together in consultation to create a collaborative concept map regarding the causal attribution present in the episode. When each participant changed the collaborative concept map, the changes were immediately reflected in the concept map, and the partner could see the changes in the collaborative concept map. However, the system did not clearly notify participants of the changes, so they were only able to see the changes if they were watching. In this process, each participant could refer to the collaborative concept map and both participants' concept maps that they had separately created in the previous individual phase. They could not see each other because of their monitors and communicated orally offline. Each participant was free to talk, and the utterances of the participants could be heard clearly from each other. There were no restrictions on the order or content of their conversations. All interactions were verbal because participants could not see each other's gestures or other physical actions. Thus, when they were considering modifying a particular node in a concept map, they often specified the location at the center of the discussion by uttering a directive and modifying the node, or by reading out the node's label.

2.4. Procedure

The entire experimental procedure is shown in Figure 3. On arriving in the experiment room, the experimenter thanked the participants for their participation. Both participants briefly introduced themselves to each other. Following this procedure, a comprehension test about causal

attribution was administered as a manipulation check to ensure they did not have prior knowledge about causal attribution. In an explanation about Cmap, the experimenter provided the task instructions on how to build concept maps and informed them that the task was to draw inferences about a certain psychological phenomenon (causal attribution). Before the task, they read a text passage about causal attribution. Thereafter, they read an episode. Then, a comprehension test (pre-test) was administered. Subsequently, in the individual phase (10 minutes), the participants engaged in the task of applying the causal attribution of success and failure to the episode. In this study, dyads were asked to build concept maps about causal attribution and create them individually. Finally, in the collaboration phase (15 minutes), the participants collaboratively created a concept map with reference to their individual maps. Finally, another comprehension test (post-test) was administered at the end.

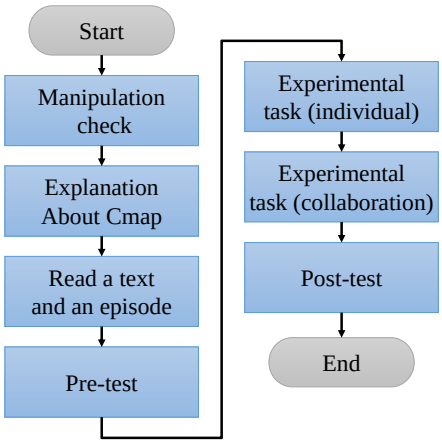

Figure 3: The entire experimental procedure. The participants performed two tasks; one was individual, and another was collaboration.

2.5. MEASUREMENT

In this experiment, participants shared the same physical space, but they could not see each other's faces and body movements. This is a situation that often occurs when individuals are using their computers to perform tasks (e.g., a scene in which a small group is teaching a collaborative learning lesson in a computer classroom). In such a situation, they can feel each other's presence in a state similar to face-to-face communication by sharing a physical space, so we thought that multimodal interaction similar to face-to-face situations might occur.

The recorded data was imported into the annotation tool Elan (https://archive.mpi.nl/tla/elan). Each data was synchronized based on the trigger signal and then used for analysis. If data preprocessing or feature extraction was required, it was performed on each original data. Once processed, the data was imported into Elan.

2.5.1. Data Segmentation

During the collaboration experimental task, facial features and gaze-direction data were acquired at approximately 30 fps. The audio data were acquired at 44,100 Hz. A moving average is a statistical method used to analyze a set of data points by creating a series of averages of different

subsets of the full data set. This is calculated by taking the sum of a set of values over a specific period and then dividing it by the number of values. In this analysis, the moving average was calculated by shifting the data by 5 seconds, with a 10-second window. We analyzed the data on the moving average. 24 participants worked for 15 minutes, resulting in 180 data per person. The total segmented data comprised 3642 passive states and 678 ACI states.

2.5.2. Utterance

The utterance is a promising clue to estimate ICAP states. We did not focus on the contents of utterances, which are difficult to analyze in real time and at high speed, limiting their application to ITS and PCA. Instead, we took the fact that the utterance was made and its duration as an index of the activity of the learner. The participants' utterances were annotated with the video start and end times by the experimenter. The number of seconds of utterance(s) included in the moving average time window was then calculated. We used the number of seconds of utterances on their own and with others as independent variables (self_utterance, other_utterance).

2.5.3. Facial Expression

In this experiment, the participants could not see each other's faces, but since they shared the same physical space and interacted with each other in the real voice, we thought that similar facial expression changes as in the face-to-face situation would appear, so we adopted the facial expression changes as an explanatory variable. The participants' facial movements were captured by video and analyzed by OpenFace, a tool that automatically calculates changes in AUs. The numerical value output by OpenFace indicated the strength of the AU and was obtained using a formula described in a "toolkit" for using the software (Baltrušaitis et al., 2016). The results of OpenFace analysis were output for each frame of the video (recorded at 30fps). From this data, the part corresponding to the moving average window was extracted, and the mean and standard deviation of the intensity of each facial expression feature in that part were calculated. These were used as independent variables (mean and SD in each time window). The following 18 types of AUs were observed among the participants: AU01: inner brow raiser; AU02: outer brow raiser; AU04: brow lowerer; AU05: upper lid raiser; AU06: cheek raiser; AU07: lid tightener; AU09: nose wrinkler; AU10: upper lip raiser; AU12: lip corner puller; AU14: dimpler; AU15: lip corner depressor; AU17: chin raiser; AU20: lip stretcher; AU23: lip tightener; AU25: lips part; AU26: jaw drop; AU28: lip suck; and AU45: blink.

2.5.4. Gaze

We considered that gaze movements contain information about what information learners are focusing on. Since it is necessary for learners to examine the information available to them at that point and consider the content of their actions in order to reach an Active state in which they actively engage their interaction partner, we thought gaze movements would help to detect readiness to move out of passivity. Gaze was acquired using the Tobii system. A monitor was placed in front of each participant to construct concept maps during the experiment, which showed three concept maps: 1) their own, 2) their partner's that had been created in the experiment's individual stage, and 3) the one on which they both were collaborating during the experiment. During the experiment, participants were free to zoom in and out and translate the screen asynchronously, making it difficult to determine which items they were looking at on the screen. In the analysis, we mapped participants' interests to the areas shown in Figure 2 (own,

partner's, or the collaborative concept maps). When looking at the own and partner's concept maps, they may be focusing on what they thought before the collaboration. When looking at the collaborative concept map, they may be focusing on what they are actually working on, what information they are changing, or considering changing the concept map through discussion. In other words, the own and partner's concept maps were referenced to identify static information, and the collaborative concept map was referenced to identify information that may change dynamically in the interaction. We used the duration of the participants' attention on this screen as independent variables (gaze_self, gaze_other, gaze_c). In addition, the distance the participant's gaze traveled on the screen was an independent variable (gaze_dist).

2.5.5. Voice

The participants' voices were captured by microphones placed in front of each participant. Although there were two participants in the same room, they rarely overlapped each other while speaking during the task. To reduce noise, only the parts of the speech uttered by each participant were isolated from the audio files. The cropped audio files were analyzed using Praat (Boersma, 2001). The features extracted in the analysis were fundamental frequency (F0), energy (Intensity), and formants (F1, F2, F3, F4). Within the moving-average window, the z-score and SD of each feature in each participant were calculated, and these were used as independent variables (mean and SD in each time window).

2.5.6. ICAP States

For the analysis, the ICAP states were annotated using the coding scheme on the interaction between the participants in video data. The ACI states were explicitly annotated and the others were regarded as the passive state. The coder watched the video and, according to the coding rules, specified the segmented section of the video that the coder considered to be Active, Constructive, or Interactive, and labeled the section accordingly. In this study, the coding scheme was developed based on Chi (2009), Chi and Wylie (2014), and Wang et al. (2015) to capture active, constructive, and interactive utterances. Table 1 shows part of the coding scheme and the relevant definitions. In addition, items of utterance about concept maps were added (e.g., utterance about concept maps; reflections of an idea on the concept map). To investigate the reliability of the coding, we conducted third-person coding based on a previous study (Schneider and Pea, 2014). The actual procedure was as follows: 1) coding rules were discussed and made by two coders, 2) two coders coded 20% of the data, 3) concordance rate was calculated, 4) two coders discussed and modified the coding rules based on the coding results, 5) whole data was coded, and 6) discordant parts of the coding were discussed and modified until the two coders agreed. In the third step, to investigate the reliability, we calculated Krippendorff's alpha coefficient. The results showed that the coder's matching rate was 0.43. The value of this indicator is not very high. So the two coders discussed why they were coding where the coding was not consistent, and then revised the coding rules and moved on to the overall coding. After the overall coding, the results of the two coders were compared, and the different parts of the coding were revised to be consistent with their agreement. Specifically, the two coders discussed which coder's coding result was appropriate, based on ICAP theory, for the part where the coding result was inconsistent. The coding rules were then updated if necessary. When the coding rules were updated, the overall coding results were checked to see if they were consistent with the updated coding rules. The most inconsistent part of the coding was the decision of whether

to be active, constructive, or interactive. This consensus-building process took approximately nine hours. These allowed us to treat the coding as reasonably reliable for the purposes of this analysis.

Table 1: Coding scheme about ICAP.

Item and definition	Example
Active: Repeat/Paraphrase The learners simply repeat the text or conversation.	"Peter said that my score was bad."
Constructive: Justify or provide reasons: The learners propose one's own idea or hypothesis.	"I think the effort is internal because of himself."
Interactive: Reflect idea on concept map: The learners agree with partner and reflect the result or understanding of partner's concept map.	After justifying or providing reasons of Constructive, "Okay, I will write effort in internal."

3. Analysis

In this experiment, we obtained nonverbal multimodal information and ICAP state data in human-human collaboration. Based on this data, we investigated cues to estimate a learner's state. Accordingly, we conducted two analyses in this study using the following: Sequential pattern mining (SPM) to obtain cues for predicting the participants' states in the near future, and Support Vector Machine (SVM) to try to classify using the obtained cues to estimate the participants' states among many variables. The participants' ICAP state annotated by the experimenter was the dependent variable, and the analyses were conducted using the abovementioned independent variables.

3.1. Sequential Pattern Mining

Pattern mining has become popular owing to its application in many domains. Some pattern mining techniques, such as frequent itemset mining (Zaki, 2000) and associative rule mining (Agrawal et al., 1994), are intended for data analysis that does not consider the order of events. Therefore, when these pattern mining techniques are applied to data that has temporal or ordinal information, this information is ignored. To solve this problem, sequential pattern mining (SPM) (Agrawal and Srikant, 1995) has been proposed. SPM is an analysis method that extends association rule mining to sequential data. SPM considers the relationship between the positions of occurrences of each item in transactional data. SPM consists of finding interesting subsequences in a set of sequences, and the interestingness of the subsequences can be measured by various criteria, such as frequency of occurrence, length, or profit. SPM has many practical applications because the data is naturally encoded as a sequence of symbols in many fields (Fournier-Viger et al., 2008; Ziebarth et al., 2015; Fournier-Viger et al., 2012; Lin et al., 2007). SPM can also be applied to time series, and when discretization is performed as a preprocessing step. We used SPM to analyze the time-series patterns in the previously mentioned features.

In SPM, the subsequences appearing in a given data series can be extracted as patterns, considering the order of occurrence of multiple elements. Using this method, we attempted to find cues to infer the learners' states by extracting the sequence patterns up to just before the change to ACI states. In SPM, the set of appearing elements is denoted as $\sum = \{i_1, i_2, ..., i_m\}$ and i_j is called an item. At this time, the data column of a series of time series is denoted as a sequence and $S = \{< a_1, a_2, ..., a_r > | a_i \in \sum\}$. A database that collects a plurality of sequences is called a transaction. A sequence indicates the order in which items appear. Because the data acquired in the experiment is a continuous value, each item was encoded as 1 (occurrence) if it exceeded a certain threshold, and 0 (absence) otherwise. For this analysis, this threshold was taken as the mean value of each item.

3.1.1. Methods and Setting

After calculating each independent variable's moving average, it was encoded as 1 if it exceeded a certain threshold (each mean) and 0 otherwise. When we checked for similar patterns in the facial expression and voice data, we found that the F1 to F4 features matched more than 90%; thus, we removed the F2 to F4 features from this analysis. The data sequence was converted into transaction data for SPM. From the transaction data (367 sequences), the data for the 15 seconds immediately before the change to ACI states in the ICAP were extracted, and the common series patterns contained in the data were extracted. We refer to these common series patterns as "IBCS (Immediately Before the Change to ACI States) patterns" in the following. To extract the rules, we used the cspade algorithm, which is among the algorithms used to perform SPM (using the rulesSequences package in R 4.1.0). The parameters for extraction in this analysis were maxlen = 4, maxsize = 2, confidence > 0.7, lift > 1.0, and support > 0.05 (maxlen: the upper limit of the length of a sequence pattern; maxsize: the upper limit of the number of types of features that simultaneously appear in a sequence pattern; confidence: the probability of a feature appearing after a sequence pattern appears; lift: the value of confidence divided by the probability of a feature appearing; and support: the probability of occurrence of a sequence pattern).

3.1.2. Results

As a result of the analysis, we could extract 96,170 patterns as the series patterns just before the change to ACI states (these are IBCS patterns). However, there were many patterns in which only one of the features was replaced. Frequent independent variables appearing in these patterns may provide cues regarding a learner's state in the near future. Table 2 shows the top 10 patterns in descending order of the value of "lift." The "Lift" column shows the value of a pattern. The "Pattern" column shows the series patterns. For example, if the series pattern is <{Intensity_SD},{F0_SD,self_utterance}>, then {Intensity_SD} initially exceeds the threshold, and then {F0_SD,self_utterance} exceeds the threshold, indicating that the state has changed to ACI states.

Observing the overall trend in the sequential patterns, the independent variables with a high occurrence frequency (variables found in more than 5% of the rules) were: self_utterance (51776, 54%), other_utterance (51507, 54%), gaze_c (15066, 16%), gaze_dist (33688, 35%), AU04 (25486, 27%), AU06 (14238, 15%), AU07 (27530, 29%), AU12 (21989, 23%), AU14 (20737, 22%), AU25 (14511, 15%), AU26 (34203, 36%), F0_SD (10291, 11%), F1_SD (5609, 6%), and Intensity_SD (16015, 17%). The figures in parentheses are the numbers of occurrences and percentages. The top 10 patterns, shown in Table 2, include only variables related to speech

Table 2: Top 10 patterns in the high value of lift.

Pattern
{Intensity_SD}, {F0_SD,self_utterance}
{F1_SD}, {F0_SD,self_utterance}
{F1_SD,Intensity_SD}, {F0_SD,self_utterance}
{Intensity_SD}, {self_utterance}
{F1_SD}, {self_utterance}
{F1_SD,Intensity_SD}, {self_utterance}
{Intensity_SD}, {Intensity_SD,self_utterance}
{F1_SD}, {Intensity_SD,self_utterance}
{F1_SD,Intensity_SD}, {Intensity_SD,self_utterance}
{Intensity_SD}, {F1_SD,self_utterance}

(both voice and utterance); however, facial expressions and gaze variables also come into play in the identified patterns. This suggests the need to observe multimodal information, not just speech, to estimate the learners' states.

In the collaboration task in the experiment, participants could not see each other's facial expressions and body movements, but they could recognize that their partner existed in the same physical space through their real voice and motion sounds. Therefore, changes in facial expressions were likely to occur in a similar way as in face-to-face situations. However, it was unlikely that a participant's state changed to an ACI state due to the other partner's facial expressions or body movements. Participants were affected by the partner's utterances and when there was a change in the collaboration concept map. Therefore, it is reasonable that the "other_utterance" was frequently included in the sequence patterns, suggesting that a participant's ACI states were also elicited from another's utterance. Meanwhile, speaking to the partner was a remarkable clue to detect the change to the ACI state of the speaker, so it is also reasonable that "self_utterance" was frequent.

Participants were also affected by the communication partner when there was a change in the collaboration concept map. Therefore, it is also reasonable that the feature "looking at the concept map that they were collaborating on (gaze_c)" was frequently found. As features of gaze, there were two other variables: "looking at the concept map created by oneself (gaze_self)" and "looking at the concept map created by another person (gaze_other)"; however, neither of them appeared in the frequent patterns. As mentioned in the Measurement section, when they were looking at own and partner's concept maps, they were probably looking at static information that they had thought about before the collaboration, and when they were looking at the collaborative concept map, they were not only actually working, but they were probably looking at information that could change dynamically by referring to information that the other person was changing. The results of SPM suggest that when a participant's state changed to ACI states, it was more likely to refer to dynamically changing information (collaboration map) than to reference information that had already been created but had not changed (own and partner's map). Since the ACI state often involves active outreach to tasks and interaction partners, it is likely that the collaboration map has been referred to more frequently to reference information that helps to infer and engage with the partner's thoughts without having to manipulate the concept map. The feature "the distance the gaze moved on the screen (gaze_dist)" is also frequently

observed. This can be considered to be a feature when comparing various information on the concept map.

As described above, it is possible to recognize the presence in the same physical space by the real voice and the action sounds, so changes in facial expressions were likely to occur in a similar manner as in face-to-face situations. Such changes in facial expressions might reflect changes in the participant's states. Facial expression features were relatively infrequently included in the sequence patterns. Although many facial features were considered, AU04, AU06, AU07, AU12, AU14, AU25, and AU26 were the variables that often appeared before a change to ACI states. AU04 (brow lowerer), AU06 (cheek raiser), and AU07 (lid tightener) are features around the eyes, which are considered to be likely to change when a participant is pondering. AU12 (lip corner puller), and AU14 (dimpler) are considered to be likely to change when a participant shows agreement. AU25 (lips part), and AU26 (jaw drop) are features around the mouth, which tend to change when a participant is trying to talk about something. Thinking and trying to talk about something can be seen as changes that emerge when preparing for active outreach to the interaction partner. Active outreach to the interaction partner can be associated with ACI status. Therefore, it is reasonable that they are observed before a change to ACI states.

Regarding voice features, the SD was a better cue than the mean to determine whether a participant's state was likely to change to ACI states. The SD changed when they started talking from silence or when they argued with others. It is possible that we perceive these situations as changes in the loudness and height of their voice. Although the mean of F0 is a commonly used index for human-state estimation, it was not an important cue in the SPM analysis. This is because we were estimating the state in the near future, and the state before the utterance might be essential. In other words, the cues for estimating a learner's state at a certain point in time might be different from those for estimating the learner's state in the near future. In the ACI state, there may be clues to these variables when the learner's state changes to other states (for example, from Active to Constructive).

Based on these results, it is probable that some consistent behavior was observed just before the change of the participant's state to ACI state, which was considered to prepare for communication with the communication partner. The independent variables that frequently appear in the sequence patterns suggest some features that are generalizable to some extent. However, SPM can easily cause combinatorial explosions and make calculations impossible; thus, the analysis only considered data before a participant's state changed to ACI states. In the operations of the concept map, there was also a behavior to modify nodes that had been manipulated a long time ago by manipulating them again. By expanding the time window of the moving average and abstracting the state transitions of the concept map, it may be necessary to consider how to analyze multiple time granularities of the SPM and extract the context of the interaction not only immediately before but also up to a certain point in time.

3.1.3. Future State Estimation

We examined the possibility of estimating whether the participants were likely to change to ACI states than passive state in the near future based on the obtained sequence pattern immediately before the change to ACI states. Specifically, we performed the SPM described above by removing one pair from the 12 pairs of participants. The data from the removed pair were then averaged over a 10-second time window, and a moving average was calculated by shifting the time window by 5 seconds; the encoding described above was then performed. The IBCS pat-

terns extracted by SPM from the 11 pairs of data were then applied to the data of the leaved pair to predict the learners' states 5 seconds later. This was repeated for 12 pairs (leave-one-pair-out cross-validation).

The number of IBCS patterns found was tallied for each pair and the z-score was calculated. The accuracy was then calculated assuming that when the value was greater than 0, the state would change to the ACI state after 5 seconds. In this case, the accuracy for all data was 63.7% (recall of passive state: 70.1%; recall of ACI state: 29.5%). This suggests that the number of IBCS patterns found provides a cue for estimating a learner's state to some degree. Meanwhile, the data of the passive state and ACI state are unbalanced. The SPM in this analysis extracted the pattern just before the ACI state for the next 5 seconds, and assumed that if the extracted pattern did not fit well, the passive state would occur. Therefore, the large number of passive states did not cause an excess of extracted patterns. However, in the case where the passive states were continuous, it was possible that the same patterns as those immediately before the ACI state appeared. In order to compare the patterns extracted from unbalanced data, it is necessary to consider an estimation method that takes into account the occurrence probability in the passive state and the occurrence probability in the ACI state. Meanwhile, the IBCS patterns found in each of the the leave-one-pair-out data were common to some extent, suggesting that they were consistent across participants, but it was difficult to estimate learner status based on SPM cues alone.

3.2. Support Vector Machine

Using SPM analysis, we investigated cues for estimating whether a learner's state would change to a non-passive one in the near future, and we found many cues in the utterances, facial expressions, gaze, and voice. It was also suggested that the patterns discovered through SPM could provide cues for estimating a learner's near-future state.

More specifically, if the multimodal information shows signs of a change in a learner's state in the near future, it may be possible to use it as a cue to estimate the change in the learner's state. To confirm this, we used a support vector machine (SVM) to estimate a learner's state 5 seconds later from data 10 seconds earlier. The number of total data samples is 4320. The variables used were the same as those in the SPM, plus the number of z-scores (numZ) of the number of IBCS patterns that could be found from the SPM results. As the SPM extracted frequent patterns from the previous 15 seconds of the data sequence, this analysis was performed on data of approximately the same duration. The kernlab package in R 4.1.0 was used for the analysis. A linear kernel was applied and the parameters were set to default.

3.2.1. SVM Model

Variable selection was performed using the backward elimination method based on the cross-validation error of the 10-fold cross-validation. The variables remaining in the SVM model with the lowest cross-validation error were AU05, AU10, AU12, AU17, AU23, AU26, gaze_dist, Intensity_mean, F1_mean, and numZ. The results show that the variables that differed from the SPM frequent variables remained. Compared to the SPM frequent variables, those related to utterance frequency were eliminated, and only AU12 and AU26 overlapped as expression variables. These overlapping variables were variables related to mouth movements, suggesting that the characteristics of participants trying to speak something may appear in common. The gaze_dist variable remained the same as that in the SPM. This is a reasonable result because a

large distance of gaze movement suggests that the participants are trying to obtain information from many locations. In addition, among the prosody variables, none were related to the SD, and those related to the mean remained. Overall, the remaining variables contain multiple types of information, suggesting that multimodal information is an important cue for estimating a learner's state.

The final classification of a learner's state in the near future according to the remaining variables resulted in a correct response rate of 75.2% for all data (recall of passive state: 80.4%; precision of passive state: 89.1%; recall of ACI state: 47.3%; precision of ACI state: 37.5%). A 10-fold cross-validation was performed, and the cross-validation error was 24.8%, while there was no difference in the percentage of correct answers. These results suggest that the multimodal information considered in this study may be useful for estimating a learner's state.

3.2.2. Leave-one-pair-out Cross-validation

Twelve pairs participated in this study. The mean number of segments per participant judged to be in the ACI state during the experiment was 28.25 (SD 12.74 [max 55, min 6]). The number of segments in the ACI state varied greatly from pair to pair, suggesting that each individual pair had different characteristics. As the number of observations was too small to create an SVM model with each pair of data, leave-one-pair-out cross-validation was performed to analyze differences in SVM models and discrimination rates. The number of training data samples was 3960, and the number of test data samples was 360 in each leave-one-pair-out cross-validation. The variables used were the same as those in the immediately preceding subsection. Table 3 shows the formula with the smallest cross-validation error among the SVM models, each excluding a data pair. Table 4 shows the percentage of correct answers, recall of passive state, and recall of ACI state, respectively.

Table 3 shows that the interaction features of each pair are so different that removing one pair changes the features effective for discrimination. Meanwhile, it also shows that some common variables are adopted. The variables that commonly remained five or more times in the SVM model were AU04, AU05, AU15, AU17, AU20, AU23, AU25, gaze_self, gaze_dist, Intensity_mean, F1_mean and numZ. The changes in AU4 and AU5 are related to the degree of eye-opening and often change when interested. The changes in AU15 and AU17 are often seen when facial expressions express dissatisfaction. The change in AU23 is often seen when people are bored. These facial expression changes may indicate participants' feelings about collaborative tasks. gaze_self indicates that they may be confirming their thoughts, while gaze_dist indicates that they are trying to take in a lot of information. These gaze characteristics may indicate the state of the participant's thoughts. Intensity_mean indicates how loud the voice is, and F1_mean is related to how the mouth opens. These voice characteristics may indicate an attitude of outreach to the person doing the collaborative work. These variables could be used to estimate learners' states independent of participants' characteristics. In a practical situation, it is difficult for a human to check all changes in the multimodal information of each participant, and it is expected that the system will realize adaptive feedback by observing the multimodal changes. Meanwhile, variables that were never employed were self_utterance, other_utterance, AU06, F0_SD, Intensity_SD, and F1_SD. As these are variables that were employed in SPM, it is possible that the numZ variables contain information. The use of SPM-derived variables in many models suggests that not only means and standard deviations but also changes in a certain time range may be important cues for estimating learner status.

Table 3: The formula with the smallest cross-validation error among the SVM models in each pair.

Leaved pair	Formula
Pair 1	category ∼ AU05_r + AU12_r + AU15_r + AU25_r + AU45_r + gaze_self + F1_mean + numZ
Pair 2	category ∼ AU01_r + AU02_r + AU05_r + AU09_r + AU17_r + AU20_r + AU23_r + AU25_r + AU26_r + F1_mean + numZ
Pair 3	category ∼ AU01_r + AU10_r + AU14_r + AU45_r + gaze_c + gaze_self + gaze_other + Intensity_mean + F1_mean + numZ
Pair 4	category ∼ AU04_r + AU09_r + AU14_r + AU15_r + AU17_r + AU20_r + AU25_r + gaze_other + F0_mean + Intensity_mean
Pair 5	category ∼ AU05_r + AU15_r + AU20_r + gaze_other + gaze_dist + Intensity_mean + F1_mean
Pair 6	category ∼ AU05_r + AU12_r + AU20_r + AU25_r + gaze_dist + Intensity_mean + F1_mean + numZ
Pair 7	category ∼ AU02_r + AU04_r + AU05_r + AU12_r + AU17_r + AU25_r + gaze_self + gaze_dist + F1_mean
Pair 8	category ∼ AU02_r + AU10_r + AU23_r + AU25_r + AU45_r + gaze_c + gaze_self + Intensity_mean + F1_mean
Pair 9	category ∼ AU04_r + AU15_r + AU17_r + AU20_r + AU23_r + gaze_c + gaze_self + gaze_dist + F1_mean + numZ
Pair 10	category ∼ AU04_r + AU09_r + AU10_r + AU14_r + AU15_r + AU20_r + AU23_r + gaze_c + F1_mean
Pair 11	category ∼ AU04_r + AU05_r + AU09_r + AU10_r + AU15_r + AU17_r + gaze_dist + F1_mean
Pair 12	category ∼ AU05_r + AU07_r + AU12_r + AU17_r + AU23_r + AU25_r + gaze_self + Intensity_mean + F1_mean + numZ

Table 4 shows that the accuracy and recall of each state in leave-one-pair-out cross-validation varies greatly depending on which pair is used as the test data. When Pairs 6 and 8 are the test data, the discrimination of passive and ACI states is relatively successful; however, when Pairs 1, 3, and 5 are the test data, the discrimination rate for the passive state decreases. When Pairs 7, 9, and 12 are the test data, the discrimination rate for the ACI states is low and unsuccessful. Pair 12 could not discriminate ACI states at all when it was the test data, which is assumed to be because this pair had the lowest number of ACI-state occurrences (7 and 6 times) and different interactions compared to the others. These findings suggest differences in the cues for estimating a learner's state for each pair, as well as a need to extract characteristics for each learner. A previous study that used LSTM to classify student behavior from video data also found that classification accuracy varied widely for each cross-validation dataset (Kasparova et al., 2021). There is a great deal of variety in the behavior of learners who participate in collaborative learning, and responding to this is a future challenge.

Table 4: The results of the percentage of correct answers, recall of passive state, and recall of ACI state in each pair.

Leaved pair	Accuracy	Recall (Passive)	Precision (Passive)	Recall (ACI)	Precision (ACI)
Pair 1	0.653	0.640	0.887	0.701	0.346
Pair 2	0.881	0.915	0.953	0.483	0.333
Pair 3	0.544	0.532	0.858	0.600	0.220
Pair 4	0.714	0.753	0.871	0.562	0.366
Pair 5	0.658	0.673	0.918	0.548	0.181
Pair 6	0.803	0.794	0.972	0.860	0.402
Pair 7	0.689	0.865	0.760	0.118	0.213
Pair 8	0.717	0.719	0.937	0.700	0.287
Pair 9	0.831	0.973	0.844	0.185	0.600
Pair 10	0.731	0.859	0.797	0.352	0.457
Pair 11	0.872	0.941	0.918	0.289	0.367
Pair 12	0.950	0.986	0.963	0.000	0.000
total	0.753	0.804	0.891	0.450	0.310

4. DISCUSSION

This study's main contribution is that it suggests the possibility of estimating a learner's state in the near future using the multimodal information expressed by the learner during collaborative learning as cues. Based on the acquired data from the experiment, an SPM analysis was conducted to search for multimodal cues to estimate from the learner's behavioral sequences whether a learner was likely to change to a state other than passive in the near future. Consequently, we could find cues evenly in speech, facial expressions, eye gaze, and voice. Furthermore, we attempted to discriminate passive states from other states (ACI states) using SVM, and the results suggest that it is possible to discriminate passive states from other states (ACI states) with an accuracy of higher than 70%. In the leave-one-pair-out cross-validation, where one individual pair was removed, an SVM model was created with data from other pairs, and the data from the removed pair were discriminated, it was possible to find variables that were included in the model relatively commonly even if the pair was removed. However, differences were also found in the accuracy, recall of passive states, and recall of ACI states, depending on which pairs were used as test data.

A comparison of the candidate cues for near-future state estimation using SPM with those using SVM shows that they tend to differ: while speech, facial expressions, eye gaze, and voice are all equally promising cues in the SPM analysis, features more related to discussion and dialog, such as the duration of speech, whether the subject is looking at a concept map, and voice height and volume variation, appear to be more promising than previously listed cues. Meanwhile, in the analysis using SVM, the variables of speech and those related to the variation of voice height and loudness (SD) were not included in the classification model, and the distance of eye movement, facial expression changes around the mouth, and mean of Intensity and F1 formant were the important variables. Based on these variables, it is possible that behaviors such as gathering a variety of information in advance and speaking clearly with the mouth relatively

wide open may reflect a learner's psychological state.

The SPM analysis showed the learners' states before they transitioned into the ACI states. They were found to be mostly in the passive state, although mentally, the learners might have already transitioned to an ACI state. In the passive state, the learners' mental states did not tend to be constant; thus, various multimodal information might have appeared. However, the features that were analyzed did not appear evenly but had a certain bias. This suggests that when changing from a passive to an ACI state, the features observed in the ACI state do not gradually increase in the passive state but rather change significantly at a certain point in time.

The novelty of this study is that it suggests the possibility of classifying the passive and ACI states of learners in the near future without depending on the contents of utterances. The ICAP framework assumes that the learner's state changes through interaction, and the Constructive state is preferable to the Active state, and the Interactive state is preferable to the Constructive state. To change the learner's state to the desired state through intervention, it is necessary to be able to infer what state the learner will change to in the near future, or how the intervention is expected to change the learner's state in the near future. If, as in previous studies, only the learner's state at a given time can be estimated, it may be difficult to support the learner to change to the desired state voluntarily, even if the intervention can induce the learner to the desired state at that time. In our study, the possibility of estimating the learner's state in the near future, even if it is only for a few seconds, was suggested, indicating that it is possible to proactively predict the learner's behavior, and to prejudge the pros and cons of an intervention and the consideration of a specific intervention method.

Schneider and Blikstein (2014) reported that learners' states could be classified into active, semi-active, or passive states by analyzing learning activities using Tangible User Interfaces, and that the passive state was significantly correlated with lower learning gains. This suggests the importance of intervening by estimating the passive state of the learners. Some previous studies have classified learners' states into active and passive states while receiving a lecture based on multimodal information (Yusuf et al., 2023), or estimated their engagement with the lectures (Zhang et al., 2020). In these studies, learners' interactions were limited, and the results obtained were not sufficient as clues to encourage learners to change from passive to active. Several previous studies have classified learners' states based on multimodal information in collaborative learning (e.g., Stewart et al. 2021; Kasparova et al. 2021). In these studies, it was possible to estimate the state at a certain time, but it was difficult to estimate the cause of the learner's state. In this study, by suggesting the possibility of estimating the future state using nonverbal information in multimodal interaction as cues, even for a short time, we were able to suggest a clue for estimating the influence from the interaction partner that can be placed in the situation where learners change to a particular state.

Meanwhile, in order to provide appropriate feedback in actual situations, decisions should often be made through direct dialogue with participants about what specific advice and interventions should be given. Since this study classifies passive and ACI states, it does not focus on speech content. However, in order to specifically classify Active, Constructive, and Interactive states and to determine the content of interventions from direct dialogue, it is necessary to refer to the content of utterances. In this study, the learner's state in the near future is estimated without speech analysis, so there is a spare time of several seconds before determining the interaction for the learner at a given time. This spare time can be applied to the processing time when speech analysis is needed to refer to the utterance content.

In order to control the learner's ICAP status in the interaction, we considered it reasonable to

quickly classify the passive and ACI states as the learner's status in the near future to determine the intervention or not, and then process to determine the intervention to move to the Active, Constructive, and Interactive states. In this study, we suggested the possibility of classifying the passive and ACI states in the near future by machine learning, which is the first part of this process, and we think it will contribute to the realization of ITS and PCA using the ICAP framework in practical situations.

Several studies have attempted to change the passive state of learners by ITS using multimodal information. For example, Wang and Yu (2024) developed an English teaching model for multimodal interaction using artificial intelligence technology. In an experiment comparing the case when this model was used with the case when it was not used, it was pointed out that the students in the control class were in a passive state. They reported that "teachers in the experimental class interacted more with students emotionally and cognitively, accepting students' positive emotions and encouraging or adopting students' viewpoints, so that students gained a sense of achievement and took the initiative to learn" (p. 15). In order to realize this kind of interaction in ITS intervention scenes, it is necessary to estimate the change in the learner's passive state, as attempted in this study, using multimodal interaction that can relatively capture the emotional aspect as a cue, and make appropriate interventions according to the learning scene. It has also been proposed to estimate emotional states during learning using ITS to provide support (e.g., McDaniel et al. 2007; D'Mello et al. 2009). However, it has also been reported that the passive emotional state is difficult to estimate because it makes it difficult to show changes in facial expressions. In the approach of this study to predict the change in the near future, it is important to capture the sign of the change rather than to estimate the state itself, and it may be possible to obtain cues for intervention by ITS, even in situations where state estimation itself is difficult.

4.1. Limitations and Future Challenges

The most major limitation is that there are large differences in the ease of estimating each pair's learning state. In order to apply to the actual situation, it is necessary to deal with the fact that the classification results of each pair vary. The main causes of such variation are the possibility that there are differences in the way learners express their states, resulting in different cues, and the possibility that learners' interaction strategies change depending on the flow of the whole discussion. With regard to the differences in how learners express their states, it is important to increase the data and include many patterns in the training data. As for the changes in learners' interaction strategies, it is possible to respond by estimating the interaction strategies from longer sequence patterns as well as the 10-second time window sequences that we are currently focusing on. The latter may be addressed by applying SPM to longer sequence patterns. It is also conceivable to classify whether a pair is in a silent interval or an utterance interval, and to analyze them separately. By separating whether participants are in a dialogue state or a thinking state, it may be possible to reduce differences in how learners represent their own states, and to improve model estimation accuracy by reducing the imbalance between passive and ACI states, especially in a dialogue state.

The second limitation, which is also important for implementing these methods, is that the acquired data may not have been sufficient to estimate the learner's state, where many patterns may exist. One of the reasons for summarizing the data as ACI states was that there were a small number of each state of Active, Constructive, and Interactive, making it difficult to

perform machine learning. In addition, human interaction in a natural state without ITS or PCA intervention will have more patterns of chronological change because there are more ways to perform tasks and more types of human speech. In the future, we would like to consider implementing ITS and PCA based on the findings obtained in this study to estimate the learner's state in a somewhat controlled interaction situation.

We are considering simultaneously measuring physiological indices as cues to estimate learners' internal states. Physiological indices such as heart-rate variability and skin-conductance responses are said to reflect the activities of the human sympathetic and parasympathetic nervous systems and provide cues to estimate human conditions such as stress. By using physiological indices together with the multimodal information used in this study, it is expected that the accuracy of ICAP-state estimation can be improved. Moreover, it is expected that not only the coding of ICAP by human observation but also its relationship with the learner's state based on physiological indices will enable the learner's state estimation to be applicable to more general situations. Regarding the voice features, we plan to acquire data from an experimental environment in which we can obtain clearer voice sounds and conduct detailed analyses. We also plan to conduct experiments in which a PCA intervenes in collaborative learning based on such state estimation. Furthermore, although a 10-second time window was adopted in this study, it was suggested that a shorter time window might be able to capture the changes more accurately. A future task is to develop a method to use data from such a short time window for near-future state estimation.

5. Conclusion

In this study, we experimentally conducted human-to-human collaborative learning using concept maps to obtain multimodal information about participants in collaborative learning to investigate cues for estimating their ICAP states in the near future. Specifically, utterances, facial expressions, eye gaze, and voice were acquired and analyzed using SPM to find cues for predicting if the learners' states would change based on the ICAP theory. We also attempted to estimate a learner's state in the near future by SVM using the sequence patterns obtained from the SPM analysis and multimodal information during learning. Subsequently, we could find several candidates that could be used for near-future-learner-state estimation. The learner-state estimation using multimodal information yielded higher than 70% accuracy. It was also suggested that capturing the characteristics of interactions in collaborative learning for each pair was necessary for more accurate estimations of a learner's state. The development of a method for inferring whether a learner's state is likely to change from the time series of interactions is a future task. Because an individual learner's ICAP state is a mental state that is difficult to observe from the outside, we are considering adding information that can provide cues to infer a person's internal state, such as physiological indicators, for more accurate estimation. In the future, we will design interventions for collaborative learning using PCAs based on state estimation for the realization of ITS. Based on the PCA proposed in Hayashi (2019a) and Hayashi (2020), we will develop a system that can detect a learner's state in real time and provide appropriate support based on the findings of this research.

REFERENCES

AGRAWAL, R. AND SRIKANT, R. 1995. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*. IEEE, 3–14.

AGRAWAL, R., SRIKANT, R., ET AL. 1994. Fast algorithms for mining association rules. In *Proceedings of 20th International Conference on Very Large Data Bases, VLDB*. Vol. 1215. Santiago, 487–499.

ALEVEN, V. A. AND KOEDINGER, K. R. 2002. An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science 26,* 2, 147–179.

ANDERSON, J. R., CORBETT, A. T., KOEDINGER, K. R., AND PELLETIER, R. 1995. Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences 4,* 2, 167–207.

BALTRUŠAITIS, T., ROBINSON, P., AND MORENCY, L.-P. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

BELENKY, D., RINGENBERG, M., OLSEN, J., ALEVEN, V., AND RUMMEL, N. 2014. Using dual eye-tracking to evaluate students' collaboration with an intelligent tutoring system for elementary-level fractions. *Grantee Submission*.

BLIKSTEIN, P. AND 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.

BODEMER, D. 2011. Tacit guidance for collaborative multimedia learning. *Computers in Human Behavior 27,* 3, 1079–1086.

BOERSMA, P. 2001. Praat, a system for doing phonetics by computer. *Glot International 5,* 9, 341–345.

CAI, Y., SHIMOJO, S., AND HAYASHI, Y. 2020. Observing facial muscles to estimate the learning state during collaborative learning: A focus on the ICAP framework. In *Proceedings of the 28th International Conference on Computers in Education (ICCE2020)*. 119–126.

CARBONELL, J. R. 1970. Ai in cai: An artificial-intelligence approach to computer-assisted instruction. *IEEE Transactions on Man-Machine Systems 11,* 4, 190–202.

CHI, M. T. 2009. Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science 1,* 1, 73–105.

CHI, M. T., ADAMS, J., BOGUSCH, E. B., BRUCHOK, C., KANG, S., LANCASTER, M., LEVY, R., LI, N., MCELDOON, K. L., STUMP, G. S., ET AL. 2018. Translating the ICAP theory of cognitive engagement into practice. *Cognitive Science 42,* 6, 1777–1832.

CHI, M. T., DE LEEUW, N., CHIU, M.-H., AND LAVANCHER, C. 1994. Eliciting self-explanations improves understanding. *Cognitive Science 18,* 3, 439–477.

CHI, M. T. AND WYLIE, R. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist 49,* 4, 219–243.

CHO, J. AND KATO, S. 2011. Detecting emotion from voice using selective bayesian pairwise classifiers. In *2011 IEEE Symposium on Computers & Informatics*. IEEE, 90–95.

DILLENBOURG, P., BAKER, M., BLAYE, A., AND O'MALLEY, C. 1996. The evolution of research on collaborative learning. In *Learning in Humans and Machines: Towards an Interdisciplinary Learning Science*, H. Spada and P. Reimann, Eds. Elsevier, 189–211.

DILLENBOURG, P. AND FISCHER, F. 2007. Computer-supported collaborative learning: The basics. *Zeitschrift für Berufs-und Wirtschaftspädagogik 21*, 111–130.

D'MELLO, S. K., CRAIG, S. D., AND GRAESSER, A. C. 2009. Multimethod assessment of affective experience and expression during deep learning. *International Journal of Learning Technology 4,* 3-4, 165–187.

EKMAN, P., FREISEN, W. V., AND ANCOLI, S. 1980. Facial signs of emotional experience. *Journal of Personality and Social Psychology 39,* 6, 1125.

EKMAN, P. AND FRIESEN, W. V. 1976. Measuring facial movement. *Environmental Psychology and Nonverbal Behavior 1,* 1, 56–75.

ENGELMANN, T. AND HESSE, F. W. 2010. How digital concept maps about the collaborators' knowledge and information influence computer-supported collaborative problem solving. *International Journal of Computer-Supported Collaborative Learning 5,* 3, 299–319.

FOURNIER-VIGER, P., GUENICHE, T., AND TSENG, V. S. 2012. Using partially-ordered sequential rules to generate more accurate sequence prediction. In *Advanced Data Mining and Applications: 8th International Conference, ADMA 2012, Nanjing, China, December 15-18, 2012. Proceedings 8*. Springer, 431–442.

FOURNIER-VIGER, P., NKAMBOU, R., AND NGUIFO, E. M. 2008. A knowledge discovery framework for learning task models from user interactions in intelligent tutoring systems. In *MICAI 2008: Advances in Artificial Intelligence*, A. Gelbukh and E. F. Morales, Eds. Springer Berlin Heidelberg, Berlin, Heidelberg, 765–778.

GOLDBERG, P., SÜMER, Ö., STÜRMER, K., WAGNER, W., GÖLLNER, R., GERJETS, P., KASNECI, E., AND TRAUTWEIN, U. 2021. Attentive or not? toward a machine learning approach to assessing students' visible engagement in classroom instruction. *Educational Psychology Review 33*, 27–49.

GRAESSER, A. AND MCNAMARA, D. 2010. Self-regulated learning in learning environments with pedagogical agents that interact in natural language. *Educational Psychologist 45,* 4, 234–244.

GWEON, G., ROSE, C., CAREY, R., AND ZAISS, Z. 2006. Providing support for adaptive scripting in an on-line collaborative learning environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, and G. Olson, Eds. CHI '06. Association for Computing Machinery, New York, NY, USA, 251–260.

HAYASHI, Y. 2018. Gaze feedback and pedagogical suggestions in collaborative learning. In *Intelligent Tutoring Systems*, R. Nkambou, R. Azevedo, and J. Vassileva, Eds. Springer International Publishing, Cham, 78–87.

HAYASHI, Y. 2019a. Multiple pedagogical conversational agents to support learner-learner collaborative learning: Effects of splitting suggestion types. *Cognitive Systems Research 54*, 246–257.

HAYASHI, Y. 2019b. Towards supporting collaborative learning with an intelligent tutoring system: Predicting learning process by using gaze and verbal information. *Bulletin of the Japanese Cognitive Science Society 26,* 3, 343–356.

HAYASHI, Y. 2020. Gaze awareness and metacognitive suggestions by a pedagogical conversational agent: an experimental investigation on interventions to support collaborative learning process and performance. *International Journal of Computer-Supported Collaborative Learning 15,* 4, 469–498.

HAYASHI, Y. AND MIWA, K. 2011. Understanding other's perspectives in conflict communication. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society 18,* 4, 569–584.

HAYASHI, Y., MIWA, K., AND MORITA, J. 2007. A laboratory study on collaborative problem solving by taking different perspectives. *Cognitive Studies 14,* 4, 604–619.

HSIAO, J.-C., CHEN, S.-K., CHEN, W., AND LIN, S. S. 2022. Developing a plugged-in class observation protocol in high-school blended stem classes: Student engagement, teacher behaviors and student-teacher interaction patterns. *Computers & Education 178*, 104403.

ILIOU, T. AND ANAGNOSTOPOULOS, C.-N. 2010. Classification on speech emotion recognition-a comparative study. *Animation 4*, 5.

JORDAN, P., RINGENBERG, M., AND HALL, B. 2006. Rapidly developing dialogue systems that support learning studies. In *Proceedings of ITS06 Workshop on Teaching with Robots, Agents, and NLP*. Springer, 1–8.

JÄRVELÄ, S., DINDAR, M., SOBOCINSKI, M., AND NGUYEN, A. 2022. Chapter 16: Multimodal research for studying collaborative learning in higher education. In *Handbook of Digital Higher Education*, R. Sharpe, S. Bennett, and T. Varga-Atkins, Eds. Edward Elgar Publishing, Cheltenham, UK, 199 – 210.

KASPAROVA, A., CELIKTUTAN, O., AND CUKUROVA, M. 2021. Inferring student engagement in collaborative problem solving from visual cues. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. ICMI '20 Companion. Association for Computing Machinery, New York, NY, USA, 177–181.

KOEDINGER, K. R., ANDERSON, J. R., HADLEY, W. H., MARK, M. A., ET AL. 1997. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education 8,* 1, 30–43.

LIN, J., KEOGH, E., WEI, L., AND LONARDI, S. 2007. Experiencing sax: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery 15*, 107–144.

MA, Y., CELEPKOLU, M., AND BOYER, K. E. 2022. Detecting impasse during collaborative problem solving with multimodal learning analytics. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. LAK22. Association for Computing Machinery, New York, NY, USA, 45–55.

MCDANIEL, M. A., ANDERSON, J. L., DERBISH, M. H., AND MORRISETTE, N. 2007. Testing the testing effect in the classroom. *European Journal of Cognitive Psychology 19,* 4-5, 494–513.

MISU, T., SUGIURA, K., KAWAHARA, T., OHTAKE, K., HORI, C., KASHIOKA, H., KAWAI, H., AND NAKAMURA, S. 2011. Modeling spoken decision support dialogue and optimization of its dialogue strategy. *ACM Transactions on Speech and Language Processing (TSLP) 7,* 3, 1–18.

MIYAKE, N. 1986. Constructive interaction and the iterative process of understanding. *Cognitive Science 10,* 2, 151–177.

MOUSAVINASAB, E., ZARIFSANAIEY, N., R. NIAKAN KALHORI, S., RAKHSHAN, M., KEIKHA, L., AND GHAZI SAEEDI, M. 2021. Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments 29,* 1, 142–163.

NASU, M. 1988. A study of weiner's attribution theory of achievement motivation. *Japanese Journal of Educational Psychology 37,* 1, 84–95.

OHMOTO, Y., SHIMOJO, S., MORITA, J., AND HAYASHI, Y. 2021. Investigating clues for estimating ICAP states based on learners' behavioural data during collaborative learning. In *Intelligent Tutoring Systems*, A. I. Cristea and C. Troussas, Eds. Springer International Publishing, Cham, 224–231.

OKADA, T. AND SIMON, H. A. 1997. Collaborative discovery in a scientific domain. *Cognitive Science 21,* 2, 109–146.

O'NEILL, I. M., HANNA, P., LIU, X., AND MCTEAR, M. F. 2003. The queen's communicator: an object-oriented dialogue manager. In *8th European Conference on Speech Communication and Technology (Eurospeech 2003)*. ISCA, 593–596.

RAUX, A., LANGNER, B., BOHUS, D., BLACK, A., AND ESKENAZI, M. 2005. Let's go public! taking a spoken dialog system to the real world. In *Proceedings of Interspeech 2005*. ISCA, 885–888.

RUMMEL, N. AND 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.

RUMMEL, N., SPADA, H., AND HAUSER, S. 2009. Learning to collaborate while being scripted or by observing a model. *International Journal of Computer-Supported Collaborative Learning 4*, 69–92.

RUMMEL, N., WEINBERGER, A., WECKER, C., FISCHER, F., MEIER, A., VOYIATZAKI, E., KAHRIMANIS, G., SPADA, H., AVOURIS, N., WALKER, E., KOEDINGER, K. R., ROSÉ, C. P., KUMAR, R., GWEON, G., WANG, Y.-C., AND JOSHI, M. 2008. New challenges in cscl: towards adaptive script support. In *Proceedings of the 8th International Conference on International Conference for the Learning Sciences - Volume 3*. ICLS'08. International Society of the Learning Sciences, 338–345.

SANGIN, M., MOLINARI, G., NÜSSLI, M.-A., AND DILLENBOURG, P. 2011. Facilitating peer knowledge modeling: Effects of a knowledge awareness tool on collaborative learning outcomes and processes. *Computers in Human Behavior 27,* 3, 1059–1067.

SCHERER, K. R., LONDON, H., AND WOLF, J. J. 1973. The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality 7,* 1, 31–44.

SCHNEIDER, B. AND BLIKSTEIN, P. 2014. Unraveling students' interaction around a tangible interface using gesture recognition. In *7th International Conference on Educational Data Mining*, J. Stamper, Z. Pardos, M. Mavrikis, and B. M. McLaren, Eds. Educational Data Mining Society, 320–323.

SCHNEIDER, B. AND 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. AND PEA, R. 2014. Toward collaboration sensing. *International Journal of Computer-Supported Collaborative Learning 9,* 4, 371–395.

SHIROUZU, H., MIYAKE, N., AND MASUKAWA, H. 2002. Cognitively active externalization for situated reflection. *Cognitive Science 26,* 4, 469–501.

SIM, K.-B., PARK, C.-H., LEE, D.-W., AND JOO, Y.-H. 2002. Emotion recognition based on frequency analysis of speech signal. *International Journal of Fuzzy Logic and Intelligent Systems 2,* 2, 122–126.

STEWART, A. E., KEIRN, Z., AND D'MELLO, S. K. 2021. Multimodal modeling of collaborative problem-solving facets in triads. *User Modeling and User-Adapted Interaction 31,* 4, 713–751.

VANLEHN, K., GRAESSER, A. C., JACKSON, G. T., JORDAN, P., OLNEY, A., AND ROSÉ, C. P. 2007. When are tutorial dialogues more effective than reading? *Cognitive Science 31,* 1, 3–62.

VOSNIADOU, S., LAWSON, M. J., BODNER, E., STEPHENSON, H., JEFFRIES, D., AND DARMAWAN, I. G. N. 2023. Using an extended ICAP-based coding guide as a framework for the analysis of classroom observations. *Teaching and Teacher Education 128*, 104133.

WALKER, E., RUMMEL, N., AND KOEDINGER, K. R. 2008. To tutor the tutor: Adaptive domain support for peer tutoring. In *Intelligent Tutoring Systems*, B. P. Woolf, E. Aïmeur, R. Nkambou, and S. Lajoie, Eds. Springer Berlin Heidelberg, Berlin, Heidelberg, 626–635.

WANG, L. AND YU, J. 2024. Research on the reform of english precision teaching in colleges and universities facilitated by artificial intelligence technology. *Applied Mathematics and Nonlinear Sciences 9,* 1.

WANG, X., YANG, D., WEN, M., KOEDINGER, K., AND ROSÉ, C. P. 2015. Investigating how student's cognitive behavior in mooc discussion forums affect learning gains. In *8th International Conference on Educational Data Mining*, O. Santos, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, and M. Desmarais, Eds. Educational Data Mining Society, 226–233.

WEINBERGER, A. AND FISCHER, F. 2006. A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education 46,* 1, 71–95.

WEINER, B. 1985. An attributional theory of achievement motivation and emotion. *Psychological Review 92,* 4, 548.

WIGGINS, B. L., EDDY, S. L., GRUNSPAN, D. Z., AND CROWE, A. J. 2017. The ICAP active learning framework predicts the learning gains observed in intensely active classroom experiences. *AERA Open 3,* 2, 2332858417708567.

YUSUF, A., NOOR, N. M., AND BELLO, S. 2023. Using multimodal learning analytics to model students' learning behavior in animated programming classroom. *Education and Information Technologies*, 1–44.

ZAKI, M. J. 2000. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering 12,* 3, 372–390.

ZHANG, Z., LI, Z., LIU, H., CAO, T., AND LIU, S. 2020. Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology. *Journal of Educational Computing Research 58,* 1, 63–86.

ZIEBARTH, S., CHOUNTA, I.-A., AND HOPPE, H. U. 2015. Resource access patterns in exam preparation activities. In *Design for Teaching and Learning in a Networked World: 10th European Conference on Technology Enhanced Learning, EC-TEL 2015*, G. Conole, T. Klobučar, C. Rensing, J. Konert, and E. Lavoué, Eds. Springer International Publishing, Cham, 497–502.