Chapter 6
Measuring Collaboration Quality Through Audio Data and Learning Analytics

Sambit Praharaj, Maren Scheffel, Marcus Specht, and Hendrik Drachsler

Abstract Collaboration is an important twenty-first-century skill. Collaboration quality detection can help to support collaboration. This chapter addresses the collaboration quality detection and measurement: (1) to define collaboration quality using audio data and unobtrusive learning analytics measures; (2) to explain the design of a sensor-based set up for automatic collaboration analytics; (3) to move toward quantifying the quality of collaboration by using this set up and show the analysis using meaningful visualizations. Furthermore, we address the challenges and issues at hand and how solutions can be built upon the work already done. To elaborate the different chapter's objectives, we use the terminology of indicators (i.e., the events) and indexes (i.e., the process) to define the components to detect collaboration quality. In one study, during collaborative brainstorming, higher was the equality (i.e., the index) of total speaking time (i.e., the indicator), lower was the dominance of each group member (in terms of total speaking time), and better was the quality of collaboration. However, quality of collaboration is dependent on the

S. Praharaj (✉)
Center for Advanced Internet Studies, Bochum, NRW, Germany

Ruhr-Universität Bochum, Bochum, NRW, Germany
e-mail: sambit.praharaj@cais-research.de

M. Scheffel
Ruhr-Universität Bochum, Bochum, NRW, Germany
e-mail: maren.scheffel@rub.de

M. Specht
Delft University of Technology, Delft, South Holland, The Netherlands
e-mail: m.m.specht@tudelft.nl

H. Drachsler
DIPF Leibniz Institute for Research and Information in Education, Frankfurt am Main, Hessen, Germany

Goethe-Universität, Frankfurt am Main, Hessen, Germany

Open University of the Netherlands, Heerlen, Limburg, The Netherlands
e-mail: h.drachsler@dipf.de

V. Kovanovic et al. (eds.), *Unobtrusive Observations of Learning in Digital Environments*, Advances in Analytics for Learning and Teaching,
https://doi.org/10.1007/978-3-031-30992-2_6

context of collaboration and the actual content of the discussion. During collaboration content analysis has been mostly on the surface level by using certain representative keywords to model different topic clusters. Therefore, we develop a sensor-based setup for automatic collaboration analytics to understand collaboration quality holistically in a learning context. Here, our aim is to understand "how" group members speak (i.e., speaking time indicator) and "what'" (i.e., the content of the conversations) group members speak to move toward collaboration quality measurement.

Keywords Collaboration analytics · Collaboration quality · Learning analytics · Group work · Technology-enhanced learning · Multimodal learning analytics

1 Introduction

Collaboration is an important twenty-first-century skill (Dede, 2010) and one of the 4Cs skill set along with critical thinking, communication, and creativity (Kivunja, 2015). Collaboration is said to occur when two or more people work toward a common goal (Dillenbourg, 1999). Most of the works in the field of learning analytics about support for collaboration have focused on analyzing remote (or online) collaboration (Jeong & Hmelo-Silver, 2010). However, with the widespread adoption of sensors (Grover et al., 2016; Kim et al., 2008), multimodal learning analytics (MMLA) (Blikstein, 2013; Di Mitri et al., 2018; Praharaj et al., 2018a) has gained prominence, thus redirecting attention to the analysis of co-located collaboration (CC) (or face-to-face collaboration) with the help of sensor technology (Grover et al., 2016; Kim et al., 2008; Praharaj et al., 2021b; Tausch et al., 2014). Moreover, sensor technology can be easily scaled up (Reilly et al., 2018) and has become affordable and reliable in the past decade (Starr et al., 2018). CC takes place in physical spaces where all group members share each other's social and epistemic space (Praharaj, 2019). Social space is composed of the non-verbal interactions (such as change in posture and specific gesture) and the non-verbal audio interactions (such as total speaking time and turn-taking). Epistemic space comprises the verbal audio interactions (such as the actual content of the conversations).

Collaboration is a complex process. "The requirement of successful collaboration is *complex, multimodal, subtle*, and learned over a lifetime. It involves *discourse, gesture, gaze, cognition, social skills, tacit practices*, etc." (Stahl et al., 2013, pp. 1–2, emphasis added). Meier et al. (2007) identified five facets of collaborative process and nine dimensions of rating collaboration quality: communication (sustaining mutual understanding, dialogue management), joint information processing (information pooling, reaching consensus), coordination (task division, time management, technical coordination), interpersonal relationship (reciprocal interaction), motivation (individual task orientation). A collaboration activity can be

called successful or not depending on the focus of the assessment of collaboration, i.e., whether collaboration is assessed as a process or as an outcome (Child & Shaw, 2015).

To measure how successful a collaborative activity is, we need to detect the quality of collaboration. Quality of CC can be detected by different indicators (i.e., the events) of collaboration such as total speaking time (Bachour et al., 2010) or eye gaze (Schneider et al., 2015). These indicators after processing and aggregation can be grouped into different indexes (i.e., the process) which act as the measurable markers of CC quality. For example, the quality of collaboration within a group can be good if there is higher equality (i.e., the index) of total speaking time (i.e., the indicator) among the group members (Bachour et al., 2010). Furthermore, different scenarios of CC such as collaborative programming (Grover et al., 2016), collaborative meetings (Kim et al., 2008; Terken & Sturm, 2010), or collaborative brainstorming (Tausch et al., 2014) each has a different set of indicators denoting the quality of collaboration. For instance, in collaborative programming relevant indicators of collaboration include pointing to the screen, grabbing the mouse from the partner, and synchrony in body posture (Grover et al., 2016); whereas in collaborative meetings gaze direction, body posture, or speaking time of group members are more relevant indicators for collaboration quality (Kim et al., 2008; Stiefelhagen & Zhu, 2002; Terken & Sturm, 2010). This difference can be attributed to the goals of the collaborative tasks and the group characteristics.

While defining indicators and indices represents the first step in measuring the quality of face-to-face collaboration, another significant challenge is the automated capturing of indicators in a scalable manner. In our work, we focus mainly on audio data, because it was the most used modality in the past studies. It can be attributed to the ease of capturing audio with a very minimalistic setup like a microphone. The CC quality has been detected from simple audio indicators of collaboration such as total speaking time and indexes like equality of total speaking time (Bachour et al., 2010; Bergstrom & Karahalios, 2007). Focus of most studies in the past was on "how group members talk" (i.e., spectral, temporal features of audio like pitch) and not "what they talk". The "what" of the conversations is more open, contrary to the "how" of the conversations in understanding what happened during collaboration (Praharaj et al., 2021b). Very few studies studied "what" group members talk about, and these studies were lab-based showing a representative overview of specific words as topic clusters (Chandrasegaran et al., 2019) instead of analyzing the richness of the content of the conversations by understanding the linkage between these words.

To overcome this, we made a starting step based on field trials to prototype, design a technical set up to collect, process, and visualize audio data automatically. The data collection took place while a board game was played among the university staff with pre-assigned roles to create awareness of the connection between learning analytics and learning design. We not only did a word-level analysis of the conversations, but also analyzed the richness of these conversations by visualizing the strength of the linkage between these words and phrases interactively. In this visualization, we used a network graph (Praharaj et al., 2021b) to visualize turn-taking

exchange between different roles along with the word-level and phrase-level analysis. This helped us to move toward automated collaboration quality detection.

Therefore, the focus of the chapter is to provide an overview of unobtrusive measures of collaboration quality (in Sect. 2) with the help of a literature review where we define the collaboration quality. Then we provide an outline of one particular method that is based on audio data. Thus, in Sect. 3, we explain the weakness of the past studies using audio data. In Sects. 4, 5, and 6, we explain our approach to move toward automated collaboration quality detection by using analytics, visualizations, and then to finally give meaningful feedback. In Sect. 7, we discuss the challenges and then in Sect. 8, we have a broader discussion, conclusion, and recommendations for future researchers in the field.

2 Defining Collaboration Quality

Collaboration quality helps us to ascertain whether a collaborative activity was successful or not. Collaboration quality is defined based on our literature review (Praharaj et al., 2021a). The broader objective of the review was to find the co-located collaboration (CC) indicators that have been detected using different modalities (such as audio, video) to understand the quality of CC.

In the first round of the analysis during the literature review, the selected publications were classified according to the sensors, indicators, and indicator types as in Fig. 6.1. One or more indicator types can be tracked using the data streams from the sensors and processing them. For instance, a microphone sensor can only track

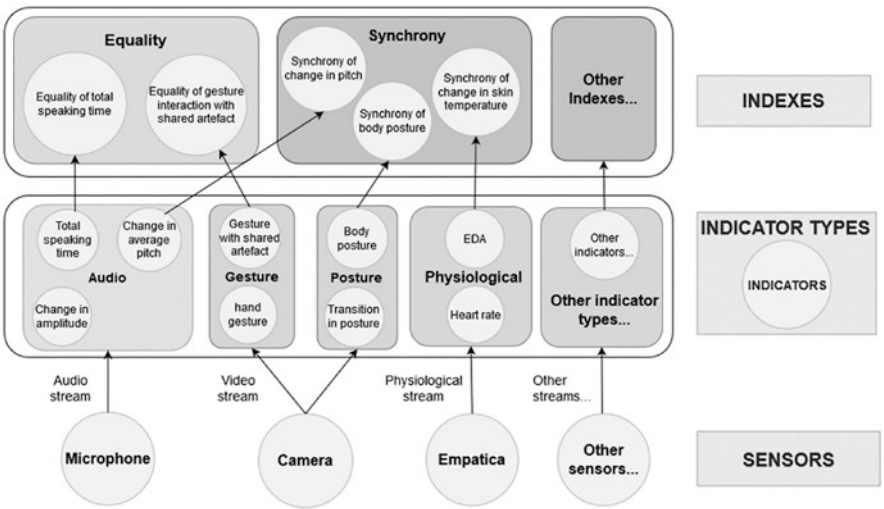


Fig. 6.1 Outline for the terminology used in the review (i.e., sensors, indicators, indicator types, and indexes) to define collaboration quality. (Reprinted from Praharaj et al. 2021a)

audio indicator type using the audio data stream whereas multiple indicator types like audio, posture, gesture, and spatial can be tracked by a Kinect (i.e., an integrated sensor which can simultaneously act as an infrared, depth, audio and video sensor). Each indicator-type cluster is composed of multiple indicators of CC detected by the sensors. For example, audio data is composed of different indicators such as pitch, amplitude, and speaking time detected by the microphone sensor.

The indicators when processed and aggregated can then be grouped to high-level indexes which define the quality of collaboration. For instance, a group which shows higher *equality* (i.e., the index) of total speaking time (i.e., the indicator) during CC has a better quality of collaboration (Bachour et al., 2010; Bergstrom & Karahalios, 2007). In the literature review, we discuss the different indicators, indicator types, and indexes of collaboration quality in-depth in more than 80 different studies with different tables which is not in the scope of this chapter. Here, we limit ourselves to the conceptual definition of collaboration quality.

But, speaking time cannot be a good indicator of collaboration across all the different scenarios of collaboration (such as collaborative programming, collaborative brainstorming, collaborative problem solving). For different scenarios, indicators of collaboration quality vary (Praharaj et al., 2018b) depending on the context. Thus, we made a scenario-driven prioritization to choose a set of indicators depending on the particular scenario of CC in the review. This formed the basis for modeling the collaboration detection framework by mapping the fundamental parameters in those scenarios onto the indicator types and indexes. There are different fundamental parameters in each scenario because of differing goals of different scenarios, team composition (such as roles and compulsory interaction with specific artifacts because of the task type), and varied group behavior (such as dominance or coupling). For example, some CC tasks already have pre-assigned roles (Hare, 1994) for each group member and in some tasks, roles emerge during collaboration (Strijbos & Weinberger, 2010). Some group members are more dominant while others are not.

Figure 6.2 shows the main outcome of the review as to how collaboration quality is detected using both the bottom-up approach (starting from the data streams of the sensors) and then the top-down approach (starting from the different scenarios of collaboration).

The mapping of these goals and parameters to the indicators and indexes to detect collaboration quality has been discussed in-depth in the literature review (Praharaj et al., 2021a) with tables. For the scope of this chapter, we will give one example from it. For example, if there is less dominance (i.e., the parameter) in the group then synchrony (i.e., the index) in body posture (i.e., the indicator) is high and the quality of collaboration is good which basically means that not one member is actively changing the posture to do the task, but everyone is actively or passively contributing to it (Kim et al., 2008).

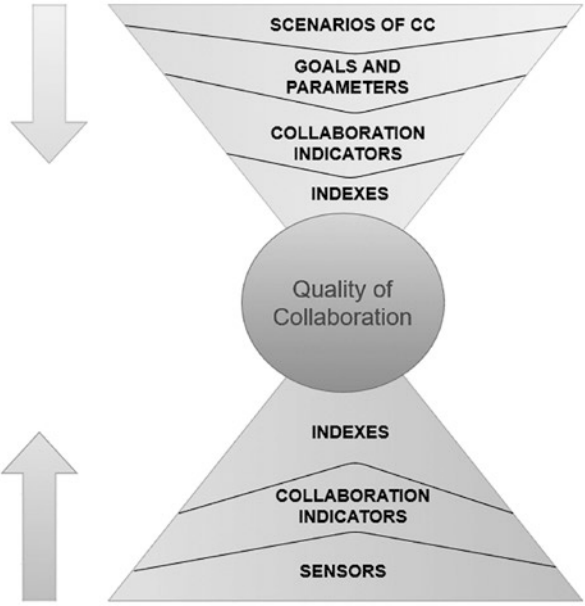


Fig. 6.2 CC quality detection using both bottom-up and top-down approach. (Reprinted from Praharaj et al. 2021a)

3 Background

We narrow our focus on group audio indicator type to detect collaboration quality only because of the abundant availability and ease of audio data collection (Praharaj et al., 2021a). Apart from the majority of studies focusing on the analysis of *how* group members speak (for instance, speaker-based indicators like the intensity, pitch, and jitter were used to detect collaboration quality among working pairs (Lubold & Pon-Barry, 2014)), very few studies used the *what* (or the content) of the audio for the analysis of CC quality.

For example, the "talk traces" (Chandrasegaran et al., 2019) and "meeter" (Huber et al., 2019) studies analyzed the content of the conversation. In the "talk traces" study, Chandrasegaran et al. (2019) did topic modeling during the meeting and then showed the topic clusters as visualization feedback by comparing with the meeting agenda. Furthermore, topic modeling is based on a collection of representative keywords which barely scratches the surface. It does not show the proper connection between these words and the rest of the conversation, which can lead to the loss of the holistic meaning of the conversations and a possible under-representation of certain topics.

The "meeter" study (Huber et al., 2019) classified the dialogues of the group members based on a lab study to measure information sharing and shared understanding while generating ideas. The collaborative task was based on three

open-ended fixed topics where group members needed to brainstorm and share their ideas in a short session of 10 min. Their performance (or the quality of collaboration) was measured based on the number of ideas they wrote down on the cards, which was quality controlled before counting the total ideas to remove bad ideas. They did not find significant effects of information sharing and shared understanding on the quality of collaboration. Therefore, the studies analyzing the content of the conversations were too abstract and mostly lab-based. To overcome these limitations, we conducted field trials to build a technical setup and then prototyped it in real-world settings to move toward automated collaboration analytics from group speech data.

Table 6.1 shows an overview of the indicators of CC and their operationalization using the group audio data in some past studies. CC takes place in physical spaces

Table 6.1 Indicators of CC and their operationalization of collaboration quality

Parameters	Indicators	Operationalizing collaboration quality	Space tracked	References
Roles (leader and follower)	Topics covered (topics are detected from keyword clusters and phrases)	Topical closeness to meeting agenda, role-based usage of keywords	Epistemic	Chandrasegaran et al. (2019) and Praharaj et al. (2021b)
Dominance	Total speaking time	Higher equality of total speaking time means less dominance and higher quality of collaboration	Social	Kim et al. (2008), Bachour et al. (2010), Bergstrom and Karahalios (2007), Praharaj et al. (2019)
Active participation	Turn-taking frequency	Frequent turn-taking changes mean higher active participation and better quality of collaboration	Social	Kim et al. (2015)
Expertise	Overlapped speech	Overlap in speech is an indicator of constructive problem solving, expertise, and good CC quality	Social	Zhou et al. (2014) and Oviatt et al. (2015)
Rapport	Synchrony in rise and fall of average pitch	Higher synchrony in rise or fall of average pitch indicates higher rapport and CC quality	Social	Lubold and Pon-Barry (2014)
Knowledge co-construction	Knowledge convergence (i.e., the amount of shared knowledge in the group), cognitive convergence	Increase in convergence (i.e., increase in shared knowledge) implies increase in CC quality	Epistemic	Jeong and Chi (2007) and Teasley et al. (2008)

Adapted from Praharaj (2022)

at the intersection of the group members' social and epistemic space (Praharaj, 2019). The *social* space consists of *how* group members speak, and the *epistemic* space consists of *what* they speak.

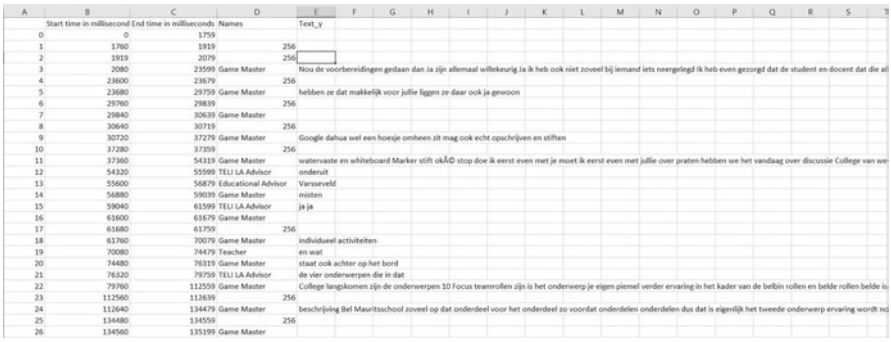
4 Automated Collaboration Analytics

To overcome the challenges, we did a field study where we looked at both the spaces to get a holistic overview of the collaboration analytics. We used the Fellowship of Learning Activity (FOLA^2) (http://www.fola2.com/, last accessed on 17 April 2023) board game where university staff with pre-assigned roles (such as teachers, all advisors (consisting of learning analytics advisor and educational advisor), learners, study coach, and game master) designed a learning activity. The main objective of this game is to create awareness of the connection between learning analytics and learning design. This game was played with different themed cards to steer the discussion in different phases for around 60–90 min in each session. In each phase, the cards had keywords related to that phase which were shown by the game master one after the other as the discussion progressed. For example, in technology phase-related discussions there were cards on interaction technologies like shakespeak and powerpoint. There were a total of 14 sessions where we recorded the audio data during the collaborative game design sessions and all these discussions were in the Dutch language. For this recording, we used clip-on microphones attached to each group member which recorded audio to the local recorder attached to those microphones.

After each game design session, these audio files were immediately transferred to the central storage space, which was the long-term storage. For the pre-processing and subsequent operations on the data, we took a copy of the files in the storage space for the pre-processing and processing unit. Here, we pre-processed and transcribed these audio files using Amber Script (https://www.amberscript.com/en/, last accessed on 28 Nov 2022). Finally, the data were processed using Natural Language Processing and analyzed to generate meaningful insights and passed on to the visualization unit to generate the visualizations. These visualizations were generated in a post hoc manner after the group meetings.

The data pre-processing, processing, analysis, and visualizations were done in Python using different openly available libraries. We pre-processed the stored audio files for each group member by extracting the timestamps from the audio file (in .wav audio file format), did speaker diarization (i.e., "who spoke when?"), and then transcribed it at the same time. Finally, we made a .csv file which contains the transcribed text, timestamps, and the roles of who spoke that text at which time. Figure 6.3 shows the data table in CSV file format after pre-processing.

This table was used to analyze the content of the conversation across sessions and role-to-role exchanges with time. We used *natural language processing* in Python for analyzing the text which includes cleaning, processing, and analyzing the text. This helped us to build the text corpus for analysis and visualizations. The following steps helped in cleaning the data:



	Start time in millisecond	End time in millisecond	Names	Text_y
0	0	1759		
1	1760	1919	256	
2	1919	2079	256	
3	2080	23599	Game Master	Nou de voorbereidingen gedaan dan Ja zijn allemaal willekeurig Ja ik heb ook niet zoveel bij iemand iets neergelegd Ik heb even gezorgd dat de student en docent dat die al
4	23600	23679	256	
5	23680	29759	Game Master	hebben ze dat makkelijk voor jullie liggen ze daar ook ja gewoon
6	29760	29839	256	
7	29840	30639	Game Master	
8	30640	30719	256	
9	30720	37279	Game Master	Google dahua wel een hoesje omheen zit mag ook echt opschrijven en stiften
10	37280	37359	256	
11	37360	54319	Game Master	watervaste en whiteboard Marker stift okÃ© stop doe ik eerst even met je moet ik eerst even met jullie over praten hebben we het vandaag over discussie College van we
12	54320	55599	TELI LA Advisor	onderuit
13	55600	56879	Educational Advisor	Varsseveld
14	56880	59039	Game Master	misten
15	59040	61599	TELI LA Advisor	ja ja
16	61600	61679	Game Master	
17	61680	61759	256	
18	61760	70079	Game Master	individueel activiteiten
19	70080	74479	Teacher	en wat
20	74480	76319	Game Master	staat ook achter op het bord
21	76320	79759	TELI LA Advisor	de vier onderwerpen die in dat
22	79760	112559	Game Master	College langskomen zijn de onderwerpen 10 Focus teamrollen zijn is het onderwerp je eigen piemel verder ervaring in het kader van de belbin rollen en belde rollen belde is
23	112560	112639	256	
24	112640	134479	Game Master	beschrijving Bel Mauritsschool zoveel op dat onderdeel voor het onderdeel zo voordat onderdelen onderdelen dus dat is eigenlijk het tweede onderwerp ervaring wordt nc
25	134480	134559	256	
26	134560	135199	Game Master	

Fig. 6.3 The stored data table sample

- Tokenization—The process of splitting the sentences into good words or tokens. It lays the foundation for the next steps of cleansing.
- Elimination of stop words—The process of removing words that mean little; these are usually words that occur very frequently. Apart from using the libraries in Python for stop word removal, we also defined our list of contextual stop words libraries that were considered unimportant for this model.
- Lemmatization and stemming—Lemmatization and stemming convert a word into its root form. For example, for the words "running" and "runs", the stem of both words is run. Thus, after we stemmed, these words would be grouped together and retain the same meaning for the model even though they had different forms.
- Sentence segmentation—We split the unstructured spoken text into different sentences, which helped the model understand the boundaries of the long text to make it more semantically distinct.
- Vectorization—Since we cannot input plain words into a model and expect it to learn from it, we had to vectorize the words. We encoded words using high-dimensional vectors where the different dimensions of vectors represent the latent meaning of the words. Therefore, the vectorized version of words would be useful later while generating bigrams (two-word combinations appearing together), trigrams (three-word combinations appearing together), and topic modeling based on the keywords or grouping semantically similar keywords.

The processed data can be used to generate different analytics and visualizations to get insights about the collaboration processes during collaborative game design.

5 Toward Collaboration Quality Detection: From Analytics to Visualizations

First, we do an exploratory analysis and visualization on the processed text data. We use topic modeling with Latent Dirichlet Allocation and Latent Semantic Indexing and then visualize the representative keywords showing different topics in one

phase of one session where the main discussion is supposed to be about technology. Figures 6.4, 6.5, and 6.6 show an overview of the topics.

Topic 1 dealt with the use of different types of interaction technology as discussed in this phase. These were mainly evident from the words: "technologie", "shakespeak", "sendstep", and "smart". These technologies were to be used by the teacher while interacting with the learner, which was evident from the word "docent", which means "teacher" in English. On examining further, the advisors (supposed to discuss technology and learning analytics) had a higher probabilistic likelihood of getting topic 1. Topic 2 refers to the use of moodle for assignments, making a photo of the post-its using the phone. This topic cluster also captured bad ("slecht") teams, ideas, and overview roles ("rol") per student. The last topical

Fig. 6.4 Topic 1: Interaction technologies

Fig. 6.5 Topic 2: Using moodle for assignments

Fig. 6.6 Topic 3: Using red cards on technology

cluster, Topic 3, focused on the use of red cards ("rod", "kaart") (or cards supposed to be used to discuss technology) and learning technology ("leertechnologie"). Then we observe the role-based bigrams and trigrams to find the interesting discussions temporally in each session. The details of the bigrams and trigrams discovered can be found in Praharaj et al. (2021b).

To do an in-depth holistic analysis of collaboration quality, we analyze both the social and epistemic space. First, we visualize the *total speaking time* and *turn-taking* from the social space and then we visualize the *content of the conversations* from the epistemic space as in Praharaj et al. (2021b). For visualizing the social space, we take the help of a node-edge network graph where each node shows a group member with a certain role and the edge shows the turn-taking between the members as in Figs. 6.7 and 6.8. The size of the node is proportional to the total speaking time of that role and the thickness of the edges is proportional to the number of turn-taking exchanges between the roles. This can help us to understand the dominant role-role exchanges temporally so that we know how the conversation patterns evolve with time.

Then, it will be interesting to visualize the epistemic space as to why certain roles have more turn-taking and dominate the conversation. Is it collaborative task-related discussion or is it clarification about the role-based tasks? To understand this further we first visualize the epistemic space to show the role-based usage of frequently used keywords during collaboration temporally. Figures 6.9 and 6.10 show the role-based usage of frequently uttered keywords in the first 20 and 30 min of the first session respectively. This helps us to understand how the usage of specific content-related or unrelated keywords is used by different roles and how it changes with time.

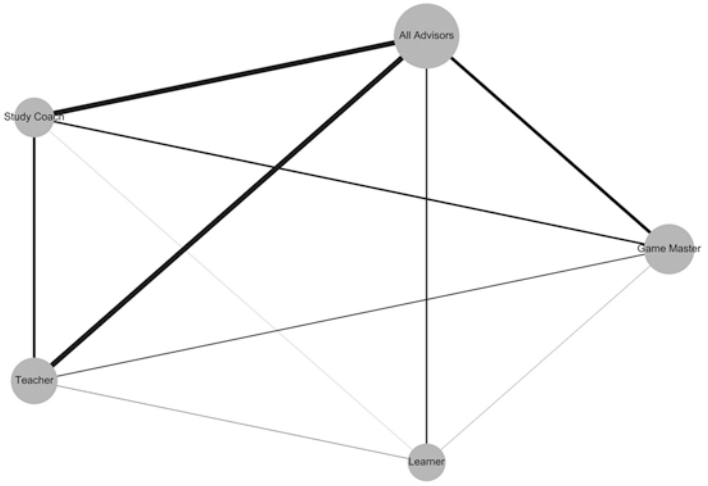


Fig. 6.7 First 20 min of social space in the first session. (Adapted from Praharaj et al. 2022)

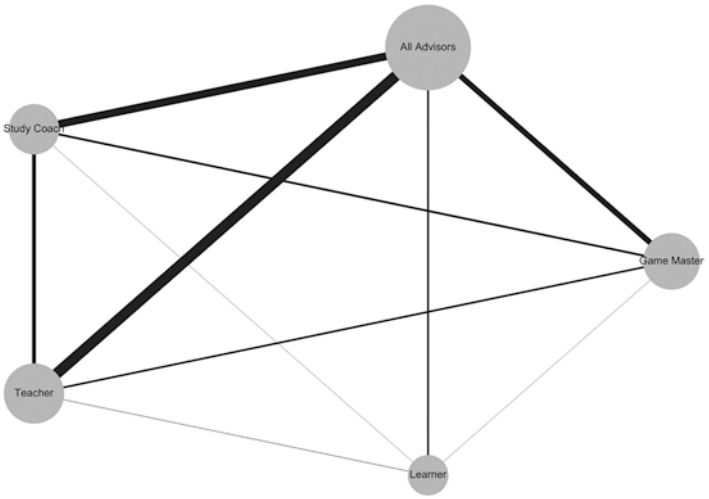


Fig. 6.8 First 30 min of social space in the first session. (Adapted from Praharaj et al. 2022)

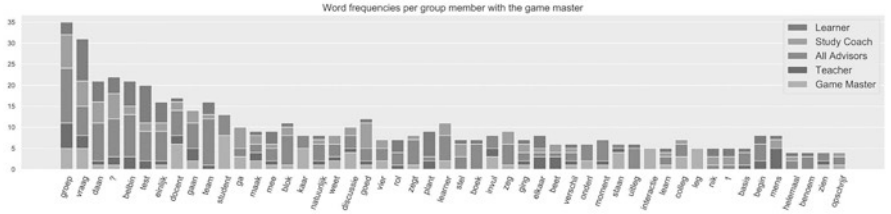


Fig. 6.9 Top 50-word utterance frequency in the first session in the first 20 min with roles. (Adapted from Praharaj et al., 2022)

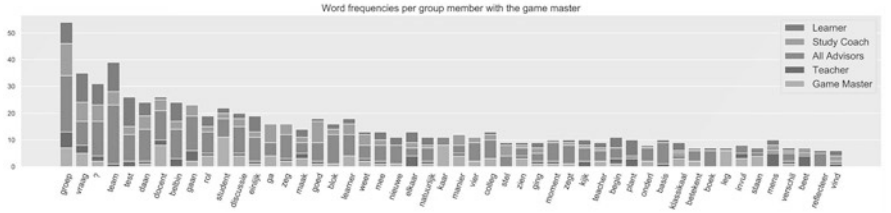


Fig. 6.10 Top 50-word utterance frequency in the first session in the first 30 min with roles. (Adapted from Praharaj et al., 2022)

Furthermore, we used the concept of knowledge convergence to quantify the quality of collaboration, i.e., how the shared knowledge among the group members (with different roles) changes as measured by the usage of different keywords with time. For instance, in Fig. 6.11, "team", a context-relevant keyword isn't spoken by the teacher in the first 10 min of the conversation but then in the next 10 min, i.e., in

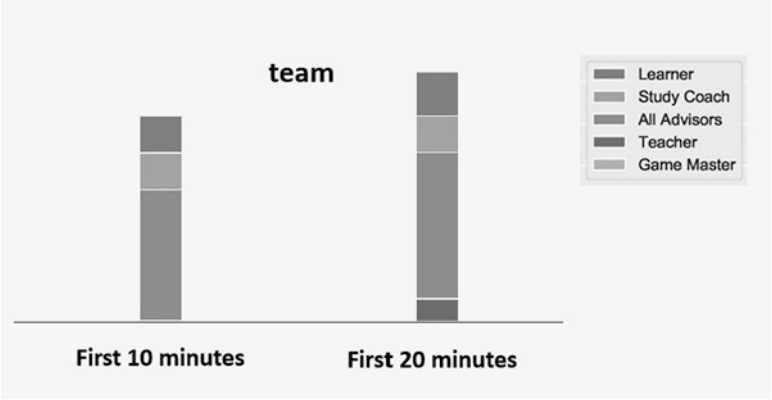


Fig. 6.11 Knowledge convergence example

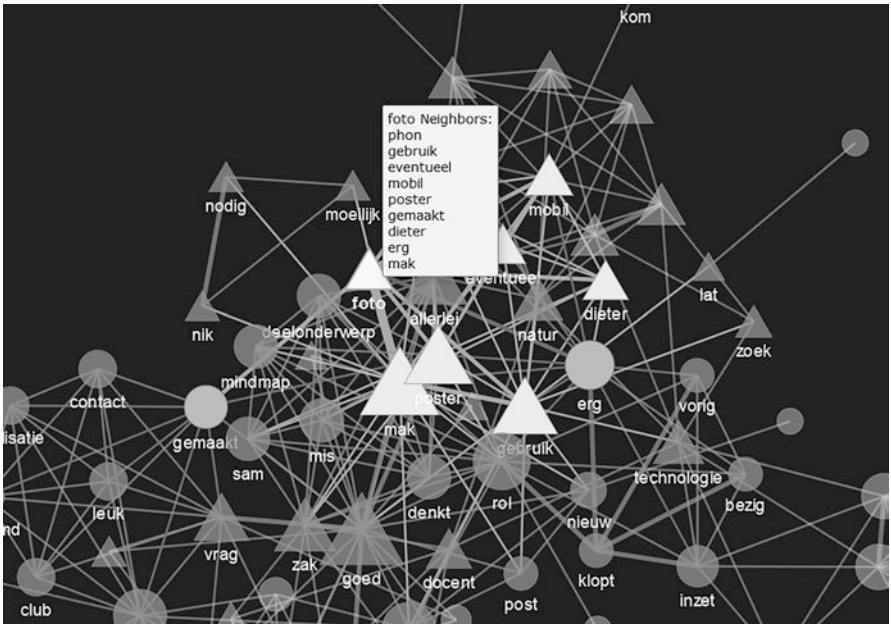


Fig. 6.12 Zoomed-in network graph highlighting a node of the advisor in rectangles and rest others in circles in technology phase of a session. (Adapted from Praharaj et al. 2022)

the first 20 min, the teacher also becomes part of the shared knowledge space of the team keyword. This signals an increase in shared context-relevant keyword knowledge convergence and thereby an increase in the quality of collaboration.

Moving from keywords to the phrases, we visualized how different words co-occur in a sentence using the network graph as in Fig. 6.12. This figure shows a zoomed-in version of the advisor role among other roles with different shape and

color. The color and shape of the node helps in the distinction of roles. The neighbors of each node (or in other words which words co-occur with each other) are shown on hovering the mouse over the node. Similarly, the strength of the words that co-occur (shown by the thickness of the edge) is also shown when we hover the mouse over the edges. The frequency of the words is proportional to the node size. This graph helps us to understand the different contextual keywords, how often they have been used, what they are associated with strongly and weakly (measured by on the edge strength of the nodes). For example, the advisor uses the words technology, mobile and photo which is associated with the use of a camera to take pictures of posters using mobile phone.

To analyze the network graph in depth, we looked at different centrality measures such as the betweenness centrality (BC) and eigenvector centrality (EC) of these words. Betweenness centrality shows how often a node (or keyword) acts as a bridge node, that is the number of times a node lies on the shortest path between other nodes. This means that keywords with high betweenness centrality are more important for the overall discussion, as they are more central in the network of keywords. Eigenvector centrality indicates the influence of a node. Therefore, a node with a high eigenvector centrality score must be connected to many other nodes who themselves have high scores. For example, in the technology discussion phase of the first session, frequency wise four words in decreasing order were "good", "make", "moodle", and "use". But, based on BC, the key terms were "good", "team", "use", and "technology", and based on EC, the key terms were "make", "poster", "good", and "role". So, this example shows that centrality measures can elevate the ranking of even less frequently used words (i.e., "team", "technology", and "role" in this example) in that context.

Figure 6.13 provides a holistic overview of the collaboration from group audio data. It shows the dashboard highlighting a node for all advisors in the technology

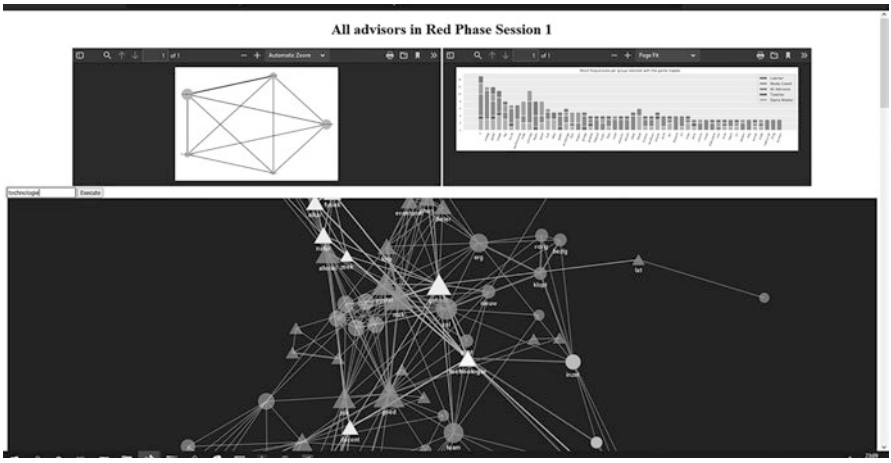


Fig. 6.13 Screenshot of the dashboard with social and epistemic components. (Adapted from Praharaj et al. 2022)

(or red) discussion phase in session 1. It has four main parts. The social space is shown by the role network graph. The high-level overview of the epistemic space is shown by the bar graph which shows role-based usage of the keywords. The colorful network graph shows the interaction of a particular role in one phase of a session. Finally, the search bar which helps to search and highlight a specific node (which is also possible on clicking on that node). Now we have different views for each phase and session with each view showing the conversation of one role in the whole conversation network graph. This will make it easier to compare two roles' conversation patterns when they are seen side by side. This dashboard is scalable, dynamic, and interactive.

6 From Visualizations to Meaningful Feedback

We will build a generic dashboard (taking help of the dashboard prototype) to quantify collaboration quality based on different collaboration indicators in the social and epistemic space with different visualizations. This dashboard will be useful to show how each role interacted during the collaboration task temporally, who was dominating the task. Now, the important question is: "Who would use it and why?". This question will be answered by understanding the needs of the dashboard design.

The design of the dashboard will be driven by the temporal needs (i.e., whether updated in real-time every few minutes or shown as a summary at the end of collaboration) and the stakeholders (teacher or task moderator or the group members themselves) who will be using it.

To address the temporal needs, we need to first differentiate what can be shown as immediate formative feedback and what can be shown as summative feedback at the end of collaboration. To this end, we need to do a qualitative study by interviewing different stakeholders to identify the user requirements. This will give us an idea as to what type of feedback is relevant for which stakeholder group and can be shown to them accordingly. For instance, this type of dashboard for a teacher (as the stakeholder) could be useful to determine scaffolding strategies during collaboration and also planning the collaboration sessions. The pedagogical meaning should be clear for the teacher to act as meaningful feedback. Is it relevant to show continuously who is dominating based on the speaking time and turn-taking or is it relevant to show certain triggers for the teacher to act like suppose when group members are confused or spending too much time in off-topic discussions? For the group members (as a stakeholder), it can be a useful tool to self-reflect (when the feedback is like a mirror) and adapt their collaboration accordingly. It might also be a more advanced version of AI-driven feedback which prompts the group members to act or behave in a certain way to enhance their collaboration.

These are some of the questions that need to be taken care of when customizing the dashboard for different stakeholders. Based on that we can also do design enhancements and modifications in the dashboard using different visualization filters to capture and compare temporal role-based snapshots.

7 Challenges

First, there are theoretical challenges. In some studies, indicators are used directly to understand the quality of collaboration without aggregating them to indexes or understanding how they contribute to collaboration quality. For example, silence has been used as an indicator of collaboration quality without understanding if more or less silence is good for the quality of collaboration. In those examples, silence was used as a feature for machine learning classifiers along with other indicators of collaboration to compute the quality of collaboration. Therefore, operationalization of the indexes to determine CC quality suffers from coding complexity even though many exist on a theoretical level (such as mutual understanding, information pooling, and others as in Meier et al. (2007)). So, there needs to be more adoption of these indexes to bring them into practice to test their strengths and limitations to understand the quality of collaboration.

Next, technical challenges are the degree of automation and the accuracy of speech to text transcription. There are challenges in processing and analyzing the data, which are largely dependent on the input (i.e., the transcribed data). The unstructured text data obtained from audio are much different than the data obtained from any online forums. Therefore, unstructured text data contains much noise, which to some extent can be structured by sentence segmentation. However, sentence segmentation working on only spoken text without punctuation marks or delimiters can cause sentence boundary detection problems. Another challenge in text processing is correcting wrongly transcribed names. For example, "moodle" was wrongly transcribed to "moeder", and we had to manually fix this in the corpus. Therefore, when studies are in-the-wild without a controlled lab environment, then there are more chances for natural, unstructured conversations, which will need cleaning and structuring before analysis can yield meaningful results.

Moreover, the stop word corpus available to the algorithm did not remove all the contextual stop words that were not relevant for this discussion. We also needed to manually remove some contextual stop words like some action verbs depending on their importance in our context by building a contextual stop word library. When we lemmatized and stemmed the words, then the lemmatizer for Dutch text was not accurate enough because of its lesser usage and popularity compared to English. Therefore, we needed to search for local libraries to correct it with some manual intervention.

It is challenging to fully automate the setup. We needed the help of a human to pre-process to some extent for cleaning the corpus, the sanity checks on the names transcribed and to make sense of the visualizations with the help of annotations. Although we are advancing toward automatic collaboration analytics, we are still in an advanced semi-automated phase and need to reduce the dependence on humans in the future.

When constructing the network graph, we quickly run into hairball problems when the graph is filled with many nodes and edges with time. It becomes very difficult to clearly distinguish individual nodes. This can be addressed while designing

in the future particularly by using temporal sliders and showing the relevant contextual keywords or words that co-occur above a certain range.

8 Discussion and Conclusion

The literature review gives an overview of unobtrusive measures of collaboration quality and helps to define the quality of collaboration as an event-process conceptual framework. Here, indicators are the events and the indexes which are obtained by processing and aggregating the indicators can be considered as the process. The indicators of collaboration quality are dependent on the scenario of collaboration because of different collaboration task goals and group characteristics (or parameters). Thus, before starting a collaboration task, it should be very clear what are the task goals, what someone wants to measure and how. This is very essential and often overlooked before starting the collaboration task. This can make the prototyping, analytics, and visualization much easier later.

Measuring the collaboration task is complex and needs operationalization of the indicators and indexes of collaboration quality. There needs to be more operationalization of the theoretical indexes into practice. This can help other researchers who want to measure the collaboration quality. For example, there has been a lot of work on measuring "sustaining mutual understanding" with human observers but there has been no work with unobtrusive sensory measures (Praharaj et al., 2021a). It is because of the contextual nuances and difficulty in understanding the content of the conversation which indicates mutual understanding from audio.

Nevertheless, the automated collaboration analytics is in an advanced semi-automated stage and humans are needed to clean the text corpus partially and also correct some names in the transcription. Therefore, there is a need to use good-quality transcription software and contextual keyword corpus to minimize the human dependence and increase the accuracy.

We find that specific keywords utterance frequency analysis for different roles helps to understand the change in role-based conversation patterns with time. This is because the more utterances we have in a specific phase-related keyword, the more is its usage in that context and hence, more importance. The convergence patterns help us to understand how specific conversations were discussed by all roles or specific roles hence signaling an increase in the shared knowledge space (i.e., a proxy for the quality of collaboration). Combined with the social space analysis (shown as role-role interaction network graph), the holistic overview of how the conversations evolved can be obtained. This helped us to quantify the collaboration quality. So, we do not categorize whether higher or lower convergence is good or bad. We just show an approach to quantify collaboration and categorizing is up to the context of collaboration. For instance, in our study, if there is higher convergence for on-topic conversations then it is good for the quality of collaboration but higher convergence for off-topic conversations is bad for collaboration quality. As

we do not define fixed objectives before collaboration and do not conduct a lab-based study, so it is quite open to interpretation.

The combined social and epistemic space also helps to clear ambiguity in certain situations when a specific indicator does not give a clear indication about the quality of collaboration. For instance, higher turn-taking signals an increase in collaboration quality only when it is happening on task-related discussion and not on clearing confusion and clarifying about the collaborative task (Kim et al., 2015). This is clear from the epistemic space or in other words the content of the conversation. So, there is a need to do a focus shift to the epistemic space from the social space and both need to be seen side by side to get a holistic overview of who spoke "what" and "how" with whom. Audio in this sense provides a richer picture of collaboration quality in an *unobtrusive* manner. With the rise of privacy and ethical concerns, anonymized audio data can be considered a good unobtrusive measure to detect collaboration quality.

Besides, there needs to be a stakeholder participatory design where their design considerations are taken into account when designing the dashboards to increase its adoption and usage. This is essential when visualizations need to be conveyed as a story on the dashboard and data storytelling can change the narrative of collaboration quality interpretation.

To conclude, our contribution is threefold: (1) to give an overview of the unobtrusive measures of collaboration where we define the quality of collaboration, (2) to build an automatic collaboration analytics setup using the audio data, and (3) to analyze and visualize the collaboration indicators from group audio data to move toward detecting CC quality.

Acknowledgments We would like to thank our colleague, Marcel Schmitz for his contributions in building this game connecting learning analytics with learning design and helping with the data collection and analysis.

References

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

Bergstrom, T., & Karahalios, K. (2007). Conversation clock: Visualizing audio patterns in co-located groups. In *40th annual Hawaii international conference on system sciences (HICSS'07)* (pp. 78–78). IEEE. https://doi.org/10.1109/HICSS.2007.151

Blikstein, P. (2013). Multimodal learning analytics. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 102–106). https://doi.org/10.1145/2460296.2460316

Chandrasegaran, S., Bryan, C., Shidara, H., Chuang, T. Y., & Ma, K. L. (2019). Talktraces: Real-time capture and visualization of verbal content in meetings. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–14). https://doi.org/10.1145/3290605.3300807

Child, S., & Shaw, S. (2015). Collaboration in the twenty-first century: Implications for assessment. *Economics, 21*, 2008.

Dede, C. (2010). Comparing frameworks for twenty-first century skills. *Twenty-first century skills: Rethinking how students learn, 20*(2010), 51–76.

Di Mitri, D., Schneider, J., Specht, M., & Drachsler, H. (2018). From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning, 34*(4), 338–349. https://doi.org/10.1111/jcal.12288

Dillenbourg, P. (1999). What do you mean by collaborative learning? In *Collaborative-learning: Cognitive & computational approaches* (pp. 1–19). Elsevier.

Grover, S., Bienkowski, M., Tamrakar, A., Siddiquie, B., Salter, D., & Divakaran, A. (2016). Multimodal analytics to study collaborative problem solving in pair programming. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 516–517). https://doi.org/10.1145/2883851.2883877

Hare, A. P. (1994). Types of roles in small groups: A bit of history and a current perspective. *Small Group Research, 25*(3), 433–448. https://doi.org/10.1177/1046496494253005

Huber, B., Shieber, S., & Gajos, K. Z. (2019). Automatically analyzing brainstorming language behavior with meeter. In *Proceedings of the ACM on human-computer interaction*, 3 (CSCW), 1–17. https://doi.org/10.1145/3359132

Jeong, H., & Chi, M. T. (2007). Knowledge convergence and collaborative learning. *Instructional Science, 35*(4), 287–315. https://doi.org/10.1007/s11251-006-9008-z

Jeong, H., & Hmelo-Silver, C. E. (2010). An overview of CSCL methodologies. In *Proceedings of the ninth international conference on learning sciences (ICLS 2010)* (Vol. 1, pp. 921–928).

Kim, T., Chang, A., Holland, L., & Pentland, A. S. (2008). Meeting mediator: Enhancing group collaboration using sociometric feedback. In *Proceedings of the 2008 ACM conference on computer supported cooperative work* (pp. 457–466). https://doi.org/10.1145/1460563.1460636

Kim, J., Truong, K. P., Charisi, V., Zaga, C., Lohse, M., Heylen, D., & Evers, V. (2015). Vocal turn-taking patterns in groups of children performing collaborative tasks: An exploratory study. In *Sixteenth annual conference of the international speech communication association*.

Kivunja, C. (2015). Exploring the pedagogical meaning and implications of the 4Cs 'super skills' for the twenty-first century through Bruner's 5E lenses of knowledge construction to improve pedagogies of the new learning paradigm. *Creative Education, 6*(02), 224. https://doi.org/10.4236/ce.2015.62021

Lubold, N., & Pon-Barry, H. (2014). Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. In *Proceedings of the 2014 ACM workshop on multimodal learning analytics workshop and grand challenge* (pp. 5–12). https://doi.org/10.1145/2666633.2666635

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

Oviatt, S., Hang, K., Zhou, J., & Chen, F. (2015). Spoken interruptions signal productive problem solving and domain expertise in mathematics. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (pp. 311–318). https://doi.org/10.1145/2818346.2820743

Praharaj, S. (2019). Co-located collaboration analytics. In *Proceedings of the 21st international conference on multimodal interaction* (pp. 473–476). https://doi.org/10.1145/3340555.3356087

Praharaj, S. (2022). *Measuring the unmeasurable?: Towards automatic co-located collaboration analytics*. Doctoral Thesis, Open Universiteit. https://doi.org/10.13140/RG.2.2.18216.65287

Praharaj, S., Scheffel, M., Drachsler, H., & Specht, M. (2018a). MULTIFOCUS: Multimodal learning analytics for co-located collaboration understanding and support. In *Proceedings of the 13th European conference on technology enhanced learning (Doctoral consortium)*.

Praharaj, S., Scheffel, M., Drachsler, H., & Specht, M. (2018b). Multimodal analytics for real-time feedback in co-located collaboration. In *European conference on technology enhanced learning* (pp. 187–201). https://doi.org/10.1007/978-3-319-98572-5_15

Praharaj, S., Scheffel, M., Drachsler, H., & Specht, M. (2019). Group coach for co-located collaboration. In *European conference on technology enhanced learning* (pp. 732–736). https://doi.org/10.1007/978-3-030-29736-7_77

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

Praharaj, S., Scheffel, M., Schmitz, M., Specht, M., & Drachsler, H. (2021b). Towards automatic collaboration analytics for group speech data using learning analytics. *Sensors, 21*(9), 3156. https://doi.org/10.3390/s21093156

Praharaj, S., Scheffel, M., Schmitz, M., Specht, M., & Drachsler, H. (2022). Towards collaborative convergence: Quantifying collaboration quality with automated co-located collaboration analytics. In *Lak22: 12th international learning analytics and knowledge conference* (pp. 358–369). https://doi.org/10.1145/3506860.3506922

Reilly, J. M., Ravenell, M., & Schneider, B. (2018). Exploring collaboration using motion sensors and multi-modal learning analytics. In *Proceedings of International Conference on Educational Data Mining*.

Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., & Pea, R. D. (2015). 3D tangibles facilitate joint visual attention in dyads. In *Proceedings of the 11th international conference on computer supported collaborative learning* (Vol. 1, pp. 156–165).

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

Starr, E. L., Reilly, J. M., & Schneider, B. (2018). Toward using multi-modal learning analytics to support and measure collaboration in co-located dyads. In *Proceedings of the 13th international conference on learning sciences*.

Stiefelhagen, R., & Zhu, J. (2002). Head orientation and gaze direction in meetings. In *CHI'02 extended abstracts on human factors in computing systems* (pp. 858–859). https://doi.org/10.1145/506443.506634

Strijbos, J. W., & Weinberger, A. (2010). Emerging and scripted roles in computer-supported collaborative learning. *Computers in Human Behaviour, 26*(4), 491–494. https://doi.org/10.1016/j.chb.2009.08.006

Tausch, S., Hausen, D., Kosan, I., Raltchev, A., & Hussmann, H. (2014). Groupgarden: Supporting brainstorming through a metaphorical group mirror on table or wall. In *Proceedings of the eighth Nordic conference on human-computer interaction: Fun, fast, foundational* (pp. 541–550). https://doi.org/10.1145/2639189.2639215

Teasley, S., Fischer, F., Dillenbourg, P., Kapur, M., Chi, M., Weinberger, A., & Stegmann, K. (2008). Cognitive convergence in collaborative learning. In *Proceedings of the eighth international conference for the learning sciences – ICLS 2008* (Vol. 3, pp. 360–367).

Terken, J., & Sturm, J. (2010). Multimodal support for social dynamics in co-located meetings. *Personal and Ubiquitous Computing, 14*(8), 703–714. https://doi.org/10.1007/s00779-010-0284-x

Zhou, J., Hang, K., Oviatt, S., Yu, K., & Chen, F. (2014). Combining empirical and machine learning techniques to predict math expertise using pen signal features. In *Proceedings of 2014 ACM workshop on multimodal learning analytics workshop & grand challenge* (pp. 29–36). https://doi.org/10.1145/2666633.2666638