JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 12(2), 105–124. https://doi.org/10.18608/jla.2025.8625

# Advancing Multimodal Collaboration Analytics: A Scoping Review

Rachelle Esterhazy[1], Rogers Kaliisa[2], Daniel Sanchez[3], Malcolm Langford[4] and Crina Damsa[5]

**Abstract**

The advent of advanced technology has opened new horizons for studying collaborative learning, although ambiguity remains in the classification and rationale for combining modalities in multimodal collaboration analytics (MMCA). Addressing this gap is crucial for the progression of collaborative learning practices and research. This review critically examines and classifies the modalities employed in MMCA studies while elucidating the rationales for their combined use and their resulting empirical contributions to collaborative learning research. A scoping review of 36 empirical studies informs the development of a framework for classifying modalities used in MMCA. We also review the rationales underlying the use of different combinations of modalities and how MMCA literature contributes to our understanding of collaboration. The review results in a definitional framework comprising five categories: auditory, visual, physiological, kinesthetic, and tactile. The findings reveal diverse arrangements of modalities. We find that the underlying rationales for combining modalities are based on technical, practical/pedagogical, methodological, or theoretical premises, which lead to different empirical contributions. Conducting MMCA research is motivated by the need for a holistic comprehension of learner behaviours, interactions, and cognitive processes during collaboration, transcending the limitations of single modalities in isolation. These findings offer both a theoretical and practical guidepost for enhancing MMCA research and applications.

**Notes for Practice**

- This review critically examines and classifies the modalities employed in MMCA studies.
- This paper proposes a comprehensive framework to categorize and define modalities.
- The rationales for combining modalities are technical, pedagogical, methodological, and theoretical.
- Application of MMCA in practice demands both technical and research competence.
- The study offers a theoretical and practical guidepost for MMCA research and applications.

**Keywords:** Multimodal collaboration analytics (MMCA), collaborative learning, scoping review

**Submitted:** 09/08/2024 — **Accepted:** 08/04/2025 — **Published:** 10/06/2025

*Corresponding author [1]Email: rachelle.esterhazy@oslomet.no Address: Centre for the Study of Professions, Oslo Metropolitan University, Stensberggata 29, 0170 Oslo, Norway. ORCID iD: https://orcid.org/0000-0003-0494-1417*
*[2]Email: rogers.kaliisa@iped.uio.no Address: Department of Education, University of Oslo, Norway, Sem Sælands vei 7, 0371 Oslo, Norway. ORCID iD: https://orcid.org/0000-0001-6528-8517*
*[3]Email: daniel.sanchez@iped.uio.no Address: Department of Education, University of Oslo, Norway, Sem Sælands vei 7, 0371 Oslo, Norway, ORCID iD: https://orcid.org/0000-0001-9771-180X*
*[4]Email: malcolm.langford@jus.uio.no Address: Department of Public and International Law, University of Oslo, Norway, Postboks 6706, St. Olavs plass, 0130 Oslo, Norway. ORCID iD: https://orcid.org/0000-0002-5975-4320*
*[5]Email: crina.damsa@iped.uio.no Address: Department of Education, University of Oslo, Norway, Sem Sælands vei 7, 0371 Oslo, Norway. ORCID iD: https://orcid.org/0000-0001-7382-4164*

## 1. Introduction

Collaboration is a key skill for employability and is essential in efforts to address complex societal problems and develop lifelong learning opportunities (Andriessen & Baker, 2020). In the most general sense, we define collaboration as a process involving more than one person working either synchronously or asynchronously to achieve a common goal, such as solving a problem or learning something. The strength of collaboration is in enabling the efficient solving of complex or ill-structured challenges or the generation of innovation (Mao et al., 2016); however, it is also important for simpler problems such as asking for help and obtaining sufficient information. Possibly more than ever before, obtaining systematic research knowledge regarding the intricate mechanisms of collaboration and how it may be supported is crucial.

Such knowledge informs education in its efforts to understand, design, support, and guide collaboration and collaborative learning for both young and experienced learners. While advances have been made in computer-assisted collaborative learning, a large proportion of collaboration occurs in collocated learning settings in which participants share the same physical space (Blikstein & Worsley, 2016). Studying individual or team behaviour in such co-located collaborative learning settings is challenging and time consuming both for researchers and teachers aiming to provide guidance for student collaboration (see review by Kim & Yoo, 2020). This challenge is, among others, related to the complex simultaneous interplay of multiple types of behaviour during collaborative activities, which are difficult to capture in a fast and reliable manner.

In educational settings, collaborative learning is usually investigated (and often assessed) through outcome or competence measures, including peer-assessment, self-evaluation surveys, think-aloud protocols, rubrics, or reflections (Sun et al., 2020; Muukkonen et al., 2020). These methods often lack the potential to provide an integrative view of the complex collaboration processes that occur in co-located collaborative learning situations. Furthermore, due to the digitization of work and education, collaboration requires the integration of knowledge across multiple interconnected systems distributed across people and machines (Wise et al., 2021). Consequently, it is challenging for both researchers and educators to fully capture and assess collaboration in authentic learning situations, as there is commonly sparse evidence to promptly identify issues arising from a team's various actions or interactions as they unfold (Graesser et al., 2020).

Nonetheless, we are witnessing a continuous development of new technologies (e.g., sensors) and methods that enable the collection and analysis of data on student collaborative processes that are otherwise not easily observable in physical and hybrid learning environments (Schneider et al., 2021). This emerging field of research is commonly referred to as multimodal collaboration analytics (MMCA). The field is under development, and numerous multimodal approaches are being utilized without a clear understanding of their substantive contribution to the study of collaborative learning. There are also inconsistencies regarding the definition of modalities (e.g., sensory channels or modes through which humans or machines interact with the world) and an understanding of the rationales underlying the combination of different modalities. This scoping review sets the course for creating an overview of existing multimodal approaches to study collaborative learning and provides a framework to classify different modalities used in the literature and underlying rationales for their combinations.

### 1.1. Contextualizing Multimodal Collaboration Analytics

Research utilizing multimodal analytics to study collaborative processes has accelerated over the last decade and grown as a sub-field of wider multimodal learning analytics (MMLA) research (Schneider et al., 2021). In this review, we define MMCA as methods of collecting data on collaborative learning through at least two modalities and processing this data (automatically or semi-automatically) to provide insight into the collaborative learning processes or outcomes. The nature of data analyzed within MMCA ranges from low-level *logs* (e.g., clickstreams that are easily captured at a scale without observers influencing the activity), *speech-based cues* (e.g., semantic or non-lexical speech features), *non-verbal indicators* (e.g., gesture, posture, and head or hand orientation), as well as *gaze and eye interaction* (Blikstein & Worsley, 2016; Martinez-Maldonado et al., 2019). These multimodal data sources can facilitate a deeper understanding of collaboration processes in co-located settings, which in turn can provide a more informed and scalable basis for feedback, assessment, and reflection in these time-bound and time-limited learning situations (Martinez-Maldonado et al., 2021; Marlow et al., 2018).

There are various literature reviews on multimodal approaches, with a few that generally focus on MMLA and others specifically on MMCA. Regarding the first stream, one of the earliest reviews of MMLA was by Di Mitri et al. (2018), who conducted a literature review of 20 empirical studies with the aim of framing the emerging field of MMLA and a focus on the nature of data and the utilization of learning theories. Chua et al. (2019) reviewed MMLA tools and technologies to collect and analyze data on learning processes in physical settings, while Giannakos and Cukurova (2023) conducted a semi-systematic review of 25 studies to explore the role of theory in MMLA studies.

The second stream of review studies focused on multimodal approaches to study collaboration. For example, Praharaj et al. (2021) conducted a literature review of 88 MMCA studies with a specific focus on co-located collaboration modelling. In particular, they sought to identify the kind of indicators deployed to understand the quality of collaboration. The findings revealed that most of the reviewed studies use audio (e.g., total speaking time, the number of interruptions while speaking, and overlap of speech) as the key indicator for collaboration quality. The study also revealed that spatial indicators (e.g., distance between group members and synchrony in posture movements) were not regarded as indicative of collaboration quality. Furthermore, Schneider et al. (2021) reviewed 74 empirical publications that used high-frequency multimodal data sources (e.g., speech, gaze, face, body, physiological, and log data) to capture facets of small collaborative groups. The primary interest was to identify the sensor-based metrics computed from multimodal data sources, operationalization of collaboration constructs, and the use of theory to interpret multimodal metrics and outcomes. The study reported several metrics ranging from physiological synchrony, linguistic features, joint visual attention, body movement/synchronization, facial expressions, and context-specific actions captured by log files. Meanwhile, the authors found inconsistencies in the definition of sensors used in MMCA studies, the underutilization of theory in interpreting metrics, and a tendency by several researchers (55% of the reviewed studies) to rely only on one modality.

Despite these reviews of MMCA, not all key issues have been examined clearly. For example, Schneider et al. (2021) included several studies in which only data from one modality was included in the final analysis, thereby reducing the potential of understanding the challenges and potentials of integrating multiple modalities. Praharaj et al.'s (2021) detailed review of MMCA studies provided a useful overview of collaboration indicators. However, similar to Schneider et al. (2021), the review provides little insight into the ways modalities were combined in the different studies and only considered research conducted in co-located settings. Finally, Chua et al. (2019) provided relevant insights into MMCA but limited their inclusion to studies conducted in physical spaces. Considering the increasing interest in adopting hybrid approaches and combining physical and digital learning environments, we argue for the need to consider all studies that utilize MMCA regardless of the learning environment.

### 1.2. Review Aims and Research Questions

Against the backdrop of a developing research field and a relatively heterogeneous use of multimodal approaches in research on collaborative learning, we argue for the necessity of a scoping review of MMCA to reveal several important issues not addressed by previous studies. In our view, these include inconsistencies regarding the definition and categorization of modalities, ways in which modalities are combined (i.e., more than just one modality), and the rationales for the choices of such combinations when studying collaboration across different contexts. To address this knowledge gap, this scoping review study aims to serve as a foundation to a more unitary understanding of multimodality that can assist in categorizing a combination of modalities utilized in MMCA research. Moreover, we aim to provide an overview of the empirical contributions of MMCA studies and their various rationales for combining modalities to research collaborative learning. To this end, our review is guided by these three research questions:

#### RQ1: How can we define and categorize modalities that are utilized in MMCA studies?

To date, there is limited consensus in the field of MMCA regarding the precise definition and understanding of modalities. Generically, multimodality refers to combining different modalities to enhance perception and provide a more comprehensive understanding of a learning activity, behaviour, or environment. Researchers sometimes mix terminologies and occasionally refer to data from the same modality as multimodal, while other authors (e.g., Praharaj et al., 2021) categorize modalities, such as audio, as indicators. By answering this research question, we aim to provide an integrated definition of modality and clarify how modalities used in MMCA studies can be categorized according to a framework built on a set of systematic criteria.

#### RQ2: Which combinations of modalities are used in MMCA studies?

Similar to the challenge of identifying a unitary terminology, the way modalities and sensors are combined in MMCA is an area that requires exploration. Researchers and labs are testing different combinations as the field and technology evolve, which leads to considerable (and potentially informative) variations of which we have little knowledge. Thus, this second research question leads us to explore the different modalities and sensors utilized in MMCA studies and how they are brought together to investigate the same phenomenon.

#### RQ3: What are the different rationales for combining multiple modalities in MMCA studies and how do they contribute to our understanding of collaboration?

Given the sheer number of potential combinations, it is relevant to understand the rationale different researchers have for selecting a specific set of modalities. It can be assumed that the authors' rationales are driven by the need to address particular research aims. These aims, in turn, may have arisen out of different kinds of knowledge gaps. Understanding the rationale (the "why") for selecting combinations of modalities has potential to offer both 1) an overview of how the empirical insights of the different studies contribute to our understanding of collaboration and 2) a better comprehension of how different rationales may affect the choice of research focus, design, and methods in MMCA research.

## 2. Materials and Methods

Aligned with our aims of generating an overview of the field, we conducted a scoping review that enables us to assess the extent, range, and nature of research utilizing multiple modalities to study collaborative learning processes (Booth et al., 2022). Scoping reviews enable an examination of a broader area of research, seeking to identify gaps in the research knowledge base, clarify key terminologies and rationales for choices of data and methodologies, and understand the conceptual boundaries of a topic (Munn et al., 2018). This methodological approach has been reported as suitable for reviewing educational research across various domains, particularly emerging disciplines or areas of research, as is the case with MMCA. The approach to the scoping review in this paper is a five-stage methodological framework, as suggested by Arksey and O'Malley (2005). The stages are 1) identifying the research question; 2) searching for relevant studies; 3) selecting studies; 4) "charting the data"[1]; and 5) collating, summarizing, and reporting the results. We found Arksey and O'Malley's framework and guidelines

[1] Extracting relevant information of relevance to the review.

sufficiently abstract and generalizable but also clear and easy to use. In the following section, we describe how the recommended steps were followed during the review process.

### Stage 1: Identifying the Research Questions

The starting point for this study was to identify relevant research questions to be addressed in the scoping review. These are presented in the previous section and cover the definition and categorization of MMCA, combinations of modalities, and underlying rationales and premises.

### Stage 2: Identifying Relevant Studies

To answer these questions, we adopted a strategy to search for relevant MMCA literature through different sources. The primary source of literature was three electronic databases: the Education Resource Information Centre (ERIC), Association for Computing Machinery (ACM), and Institute for Electrical and Electronics Engineers (IEEE). Since most research on learning analytics is indexed in these databases, they were regarded as suitable for finding the most relevant studies. In addition, we found it valuable to employ a snowball approach and check reference lists of the studies found through database searches. This process identified 33 extra articles, strengthening our review's validity. Additional searches were conducted in Google Scholar to identify other studies that might have been missed in the three principal databases. Lastly, a hand-search of the flagship journal and conference in learning analytics (i.e., the *Journal of Learning Analytics* and the proceedings of the International Conference of Learning Analytics and Knowledge) was performed to check for additional articles that met the criteria. As to the time period, we did not set a strict start date but included all studies found until 31 January 2023, after which no additional studies were added. To find relevant studies, we created a two-dimensional keyword strings with different sets of words guided by previous MMCA reviews and the study's research questions. These were as follows: ((collaborat* OR teamwork OR team-based OR "group work") AND (educati* OR learning OR problem-solving)) AND (("learning analytics" OR "educational data mining" OR "educational data science" OR "collaboration analytics" OR "team analytics") AND (multimodal* OR multi-modal* OR modal*)). To manage the bibliographic references retrieved from the different databases, we used ZOTERO, a bibliographic reference manager, to save and sort the retrieved studies.

### Stage 3: Selecting Relevant Studies

Overall, the different search mechanisms applied in our review generated 163 references. To systematically select relevant studies, we defined inclusion/exclusion criteria to help us eliminate those that did not align with the objectives of our review. Studies had to be 1) peer-reviewed, 2) in English, 3) refer to empirical data, 4) use more than one modality in the data collection and analysis, and 5) engage with multimodal data to study aspects of collaboration or teamwork (i.e., data is collected while at least two people are interacting). On this basis, we excluded literature reviews, project plans, posters, and PhD theses, but included conference papers and short papers (on the basis that there was sufficient peer review). We also excluded studies analyzing one modality, duplicate studies published in different venues, and studies utilizing more than one modality but not focusing on collaboration. Guided by the inclusion criteria, all records were screened by two researchers. The initial 163 records were screened by reading the abstract; if that was insufficiently clear, the paper was subjected to full-text reading. This process resulted in a reduced number of articles (N=60). The next stage involved reading the full text of these 60 articles to decide which ones to include in the final review. After reading the articles in full text, 36 articles were selected, in accordance with the criteria, for inclusion in the review (see Figure 1 for a detailed flow of the selection process).

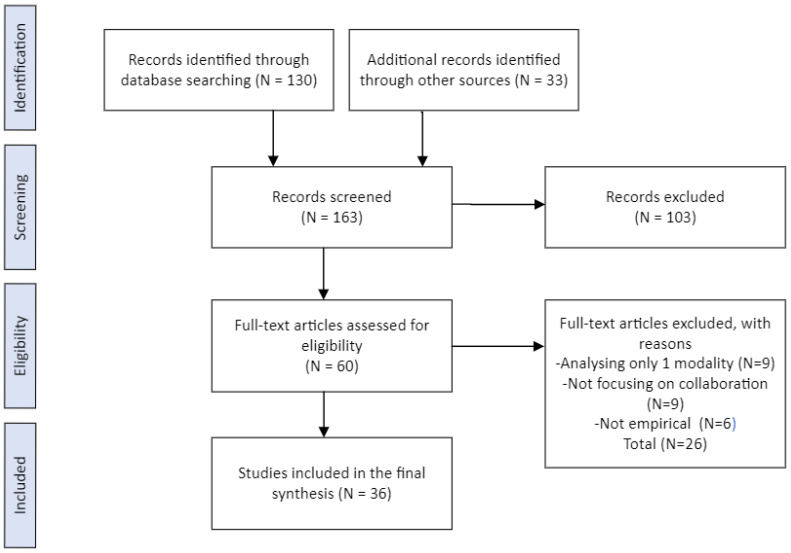


**Figure 1.** Flow diagram illustrating the scoping review process (adapted from Moher et al., 2009).

### Stage 4: Charting the Data

Charting scoping reviews refers to extracting relevant information from the identified studies based on key topics and themes of relevance to the review (Arksey & O'Malley, 2005). The charting of the data was a process that led to the selection of articles but that also fed into the development of a framework for classifying the examined studies. This is reported on in the findings section. In this review, key items of information were extracted by five researchers utilizing a uniform extraction template (see Appendix 2) that included general information regarding the study and specific information of relevance to the study's research questions. To ensure the validity and reliability of this extracted information, all studies were coded by at least two of the five authors. We used social moderation for inconsistencies and engaged in discussion until agreement was reached. To address RQ1, we analyzed the types of data collected by the authors (e.g., x-y coordinates) and what information about the participants was extracted from the different data types (e.g., spatial position or movement by participants). Based on this overview, we categorized each study into the different sensory channels or modes through which the data was captured, resulting in the definitional framework of modalities as elaborated below. To answer RQ2, we then charted the combinations of modalities in the reviewed literature using our own framework. Finally, addressing RQ3, we grouped the studies based on the author's rationales identified through our extraction template and created a synthesis of the empirical insights on collaboration that could be gained.

### Stage 5: Collating, Summarizing, and Reporting the Results

One of the key aims of a scoping review is to summarize and present an overview of all the information extracted from the reviewed studies. Using the extracted information as described in stage four, we developed an account of the findings through both quantitative and qualitative analysis. As to the first, we performed a simple quantitative analysis of the studies' demographic characteristics (e.g., year of publication, geographical spread, educational settings, and discipline of participants). This is presented in the form of charts and descriptions in the beginning of the results section, followed by a presentation of our findings organized thematically by research question.

## 3. Results

### 3.1. General Description of the Reviewed Studies

Thirty-six studies were included in the review. Learning analytics and MMCA is a broad interdisciplinary research field that encompasses various epistemologies, ontological approaches, and methodologies. This is a natural consequence of the fact that researchers using MMCA come from different backgrounds. In this section, we present a descriptive account of the geographical background, the publication year, as well as the disciplinary and educational settings of the included studies. Based on the institutional home of the lead author, we found that the US was the most represented country in our sample, with 14 contributions (see Figure 2). The research on MMCA in the US is spread both over different institutions and research groups that focus on either technology or learning. The earliest MMCA study included was conducted at the MIT Media lab (Madan et al., 2004), while the US group with the most contributions was the Learning, Innovation, and Technology lab at Harvard University (Schneider & Pea, 2015; Sinclair & Schneider, 2021; Huang et al., 2019; Reilly & Schneider, 2019). Australia had the second-highest representation overall, with nine studies affiliated with the Centre for Learning Analytics at Monash University and the team led by Roberto Martinez-Maldonado. They have been working with multimodal collaboration since 2013, with a rapid increase in the number of publications since 2017. Eight Europe-based studies were included with early explorations around 2007 in the Netherlands (Sturm et al., 2007), but no further publications until the 2017 EU project PELARS[2] (Spikol et al., 2017; Spikol et al., 2018). In addition, our sample includes MMCA research from Germany, Sweden, Estonia, and Finland. Finally, there were two studies from Japan and one from Ecuador. For a distribution of studies per year of publication, see Figure 3.

Regarding the educational settings in which MMCA is conducted, most studies take place within higher education institutions, with 26 studies dedicated to exploring MMCA in this context. This strong emphasis on understanding and enhancing collaborative practices within higher education settings is perhaps unusual given the usually strong focus on primary and secondary education in learning analytics. Secondary education was represented in seven studies, while primary education had the lowest representation with only one (Olsen et al., 2020). Two studies did not specify the empirical setting. These findings emphasize the importance of MMCA in various educational contexts and raise questions regarding whether there is a need to focus on lower levels of education, particularly given that prior experience with groups affects future participation and perceptions (Theobald et al., 2017). See Figure 4 for the distribution of studies across settings.

[2] Practice-based Experiential Learning Analytics Research and Support.

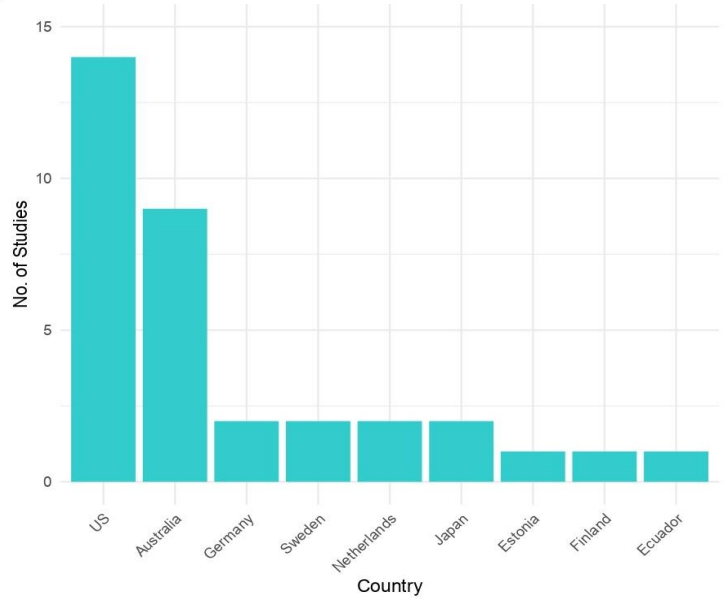


**Figure 2.** MMCA studies by country.

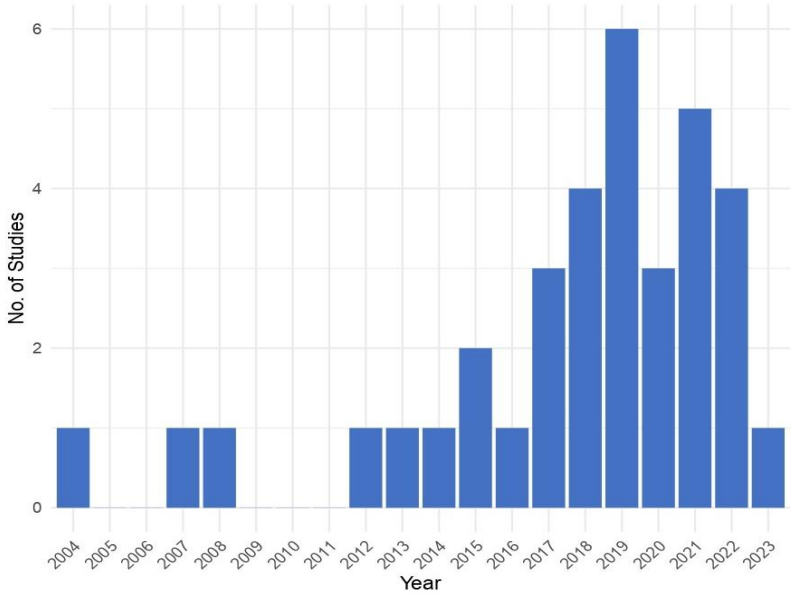


**Figure 3.** MMCA studies by publication year.

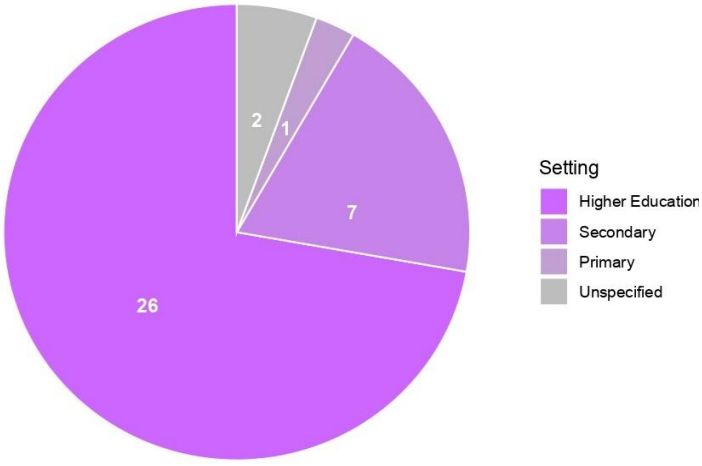


**Figure 4.** MMCA studies by setting.

After having established the characteristics of the body of literature in this review, we now address our three research questions. We first present how we can define and categorize the modalities employed in the reviewed MMCA studies (RQ1); thereafter, we describe which combinations of modalities and sensors are used in the studies (RQ2). Finally, we identify the rationales justifying the use of multiple modalities and sensors in the reviewed literature (RQ3).

### 3.2. Definitional Framework of Modalities Utilized in MMCA

One of the main objectives of this scoping review was to understand how we can define the modalities used in MMCA studies. This was motivated by the lack of consensus on what multimodality entails in a typical MMCA study. Based on systematic analysis of the empirical studies, as well as consulting methodological literature, we generated a definitional framework of the different modalities utilized in the reviewed MMCA literature. We define *modality* as the distinct sensory channels or modes through which humans (or machines) perceive and interact with the external environment. It involves the recognition and interpretation of stimuli by specific sensory organs or technical sensors, which enable the detection and processing of sensory information. Stimuli represent the external events, patterns, or objects that trigger sensory experiences (Fritzsch, 2021). These stimuli activate the corresponding sensors, initiating the conversion of sensory information into digital or neural signals that are further processed in our nervous system or within the circuits of a computer. This definition is framed in relation to constitutive elements of digital-material or immaterial nature — such as stimuli, sensors, and physiological functions — all of which are borrowed from literature on cognitive science and computer–human interaction (Fritzsch, 2021; Oviatt & Cohen, 2014; Quek et al., 2002; Turk, 2014).

It follows from this definition that a modality can be described by the relationship between the specific sensors and the corresponding stimuli that they detect. For example, an audio modality is registered after sound waves (the stimulus) stimulate a microphone or human ears (the sensor). The traditional classification of modalities in human perception are the visual modality (eyes detecting light waves), auditory modality (ears detecting sound waves), tactile modality (nerves in skin or muscles detecting pressure or temperature changes), olfactory modality (nose detecting chemical molecules), and gustatory modality (tongue detecting chemical molecules).

The way humans interact and communicate with the world is inherently multimodal (Bunt et al., 1998; Quek et al., 2002). *Multimodality* refers to the combination of different modalities, which provides a more comprehensive understanding of the environment. This integration enables humans and machines to collect information from various sources, thereby creating a potentially more nuanced perception of the world.

Most MMCA technologies are developed with the aim to imitate the way the human sensory system perceives such events, patterns, and objects — that is, stimuli — that provide us with relevant multimodal information regarding someone else's learning. For example, just like a teacher might draw inferences regarding her students' learning by perceiving certain visual and auditory cues, MMCA research may attempt to make similar inferences by using a set of technical sensors imitating the teacher's vision and hearing. In addition, new technical sensors — such as heart rate monitors or skin conductance sensors — allow us to add modalities that make visible emotional or physiological responses that would not have been accessible to the teacher's senses otherwise.

In the reviewed literature (see Table 1), we found examples of studies that collected data from sensors "imitating" three of the main sensory modalities found in human perception: sensors detecting sound waves (auditory modality), sensors detecting light waves (visual modality), and sensors detecting pressure (tactile modality). In addition, we found examples of studies using sensors detecting movement (kinesthetic modality) and sensors that detected electric signals generated by the body (physiological modality). The latter modality is unique for machine perception and cannot easily be achieved by humans without technical help. Maybe not surprisingly, we did not find examples of studies that used sensors collecting data from olfactory or gustatory stimuli.

From each of these modalities, different types of information about the learners can be extracted. In the reviewed literature, we found that the following types of information were used. In the auditory modality, sound waves emitted by student voices were used to indicate prosodic (how) or semantic (what) aspects of their speech. In the visual modality, information extracted from light waves was used to indicate learner positions or movements in the room, the direction of their gaze, their facial expressions, and their concrete actions. In the kinesthetic modality, data from accelerometers or position trackers were used to indicate learner positions or movements in the room. In the tactile modality, whenever students touched tangible user interfaces, tabletops, or keyboards, these interfaces could capture information regarding specific actions (e.g., digital logs of what learners click on in a digital application) or production of written text (e.g., learner keyboard input in an online learning environment). Finally, in the physiological modality, the electrical signals or chest movement picked up by the sensors are used to measure heart rate, breathing rate, or skin conductance, all of which are proxies for the physical arousal of the learner. See Figure 5 for an illustration of which information can be extracted from the learner by means of the five modalities identified in the literature.

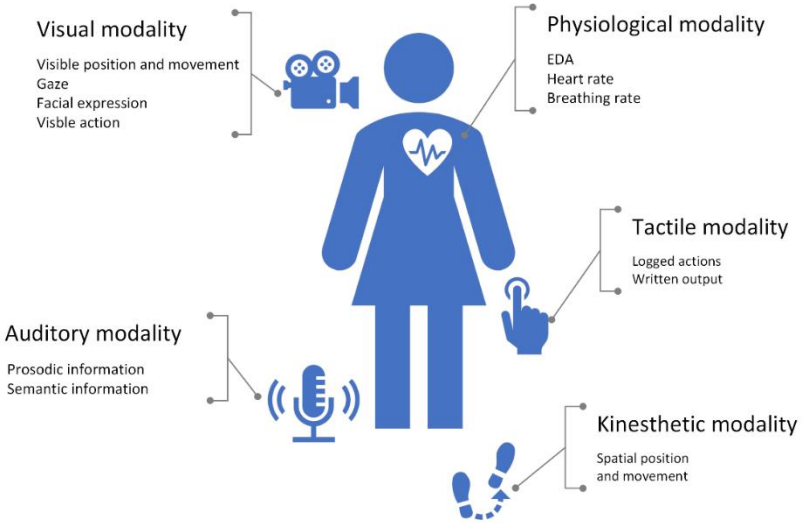


**Figure 5.** Modalities found in the literature and information extracted about the learner.

In certain cases, the same type of information on the learner can be extracted from two different modalities that use fundamentally different technologies to achieve similar results. For example, the learner's position and movement can be identified via a Kinect camera that captures the light wave patterns reflected by the learner in movement as well as via an accelerometer that registers changes in mass when the body to which it is attached is moving. Another example is learner actions that can be identified via a camera detecting specific light wave patterns reflected by the learner while performing an action or by using tangible user interfaces (e.g., tablets, interactive tabletops, high-fidelity manikins) that detect touch/pressure that indicates that the learner is performing specific actions.

In summary, the analysis of the different sensors and modalities used in the reviewed studies results in our definitional framework of modalities in MMCA literature. See Table 1 for the complete framework and specific examples from our review.

**Table 1.** Definitional Framework of Modalities in MMCA Literature

| Modality | Extracted info on the learner | Sensor/ device used | Description | Example from the literature |
|---|---|---|---|---|
| Auditory | Prosodic | Microphone | Prosodic information is registered when the microphone detects sound wave patterns that indicate aspects of speech (e.g., volume, intonation, speed). | Scherer et al. (2012) detected the "speak slope" of the audio file to measure breathy or tense voice quality. |
| | Semantic | Microphone | Semantic information is registered when the microphone detects sound wave patterns that are transformed into a written transcript and indicate speech content. | Sinclair and Schneider (2021) detected "vocabulary overlap" by calculating the ratio of shared vocabulary present in the dialogue between participants. |
| Visual | Position and movement | Kinect camera | The *visible position* or *movement* of the learner's body or body parts is registered when specialized Kinect cameras detect light wave patterns reflected by the learner in movement. | Huang et al. (2019) detected the "vertical difference in head orientation" between two learners based on Kinect coordinates. |
| | Gaze | Eye tracking camera | The *visible direction and movement of the learner's gaze* is registered when specialized eye tracking cameras detect light wave patterns reflected by the learner's moving eyes. | Reilly and Schneider (2019) detected "joint visual attention" by measuring the proportion of time both participants were looking at the same spot. |
| | Facial expression | Camera | The learner's *visible facial expression* is registered when a camera detects light wave patterns reflected by the movement of expressive facial muscles. | Ma et al. (2022) detected "facial expression" by measuring the action unit vectors from the faces captured on video. |
| | Action | Camera | The learner's *visible actions* are registered when a camera detects specific light wave patterns reflected by the learner while performing an action. The interpretation typically involves human judgment but can, in certain cases, be done automatically. | Vrzakova et al. (2020) identified student "solution attempts" by detecting changes in pixels during a screen recording of an online CPS session. |
| Kinesthetic | Position or movement | Accelerometer; Position tracker | The learner's *spatial position or movement* is registered when an accelerometer or position tracker attached to the body detects changes in coordinates. | Zhao et al. (2022) detected whether learners are in an "interactional space" when their coordinates are in close proximity to another learner. |

| | | | | |
|---|---|---|---|---|
| Tactile | Actions | Tangible user interfaces | The learner's *logged actions* are registered when tangible user interfaces (e.g., tablets, interactive tabletops, high-fidelity manikins) detect touch/pressure, indicating that the learner is performing specific actions. | Echeverria et al. (2019) detected "manikin manipulation" by logging all actions performed on a high-fidelity manikin. |
| | Written output | Keyboards<br>Digital pens | The learner's *written output* is registered when keyboards or digital pens detect touch/pressure indicating that the learner is writing something. | Chejara et al. (2020) detected the "number of characters added by each student" through a keyboard using a writing software. |
| Physiological | Skin conductance | Electrodermic activity sensor | The learner's *skin conductance* is registered when an electrodermic activity (EDA) sensor detects changes in the conductivity in the learner's skin due to increases in the activity of sweat glands. | Martinez-Maldonado et al. (2020) detected learners "physiological activation" through an EDA wristband. |
| | Respiration rate | Respiration rate sensor | The *respiration rate* is registered when a sensor detects expansion in the learner's chest. | Ouhaichi et al. (2021) detected the "respiration rate" of participants. |
| | Heartrate | Heartrate sensor | The *heartrate* is registered when a heartrate sensor detects changes in the electrical signals generated by the learner's heart. | Neubauer et al. (2016) detected the "heartrate variability" through a mobile heartrate sensor. |

### 3.3. Combinations of Modalities Utilized in MMCA Studies

The second research question aimed to explore the combinations of modalities used in MMCA studies. In line with the review focus, all studies included in this review combined at least two of the five modalities defined in the above framework. As presented in Table 2, the most common combination were two modalities (N=17), followed by three modalities (N=16), and four modalities (N=3).

**Table 2.** Overview of Studies Organized by 1) Combination of Modalities and 2) Underlying Rationales

| | **Underlying rationale** | | | | |
|---|---|---|---|---|---|
| **Combinations** | **Technical** | **Pedagogical/ Practical** | **Methodological** | **Theoretical** | **Total** |
| Auditory + Visual | Sturm et al. 2007 | Stewart et al., 2021; Vrzakova et al., 2020; Praharaj et al., 2018; Müller et al., 2018 | Schneider & Pea, 2015 | Sinclair & Schneider, 2021; Ma et al., 2022 | 8 |
| Auditory + Tactile | | Scherer et al., 2012; Viswanathan & VanLehn, 2018 | Chejara et al., 2020 | | 3 |
| Visual + Physiological | Malmberg et al., 2019 | | Noroozi et al., 2019 | | 2 |
| Auditory + Kinesthetic | | Kim et al., 2008 | Zhao et al., 2023; Zhao et al., 2022 | | 3 |
| Visual + Tactile | Spikol et al. 2017a | | | | 1 |
| Auditory + Visual + Tactile | Martinez-Maldonado et al., 2013; Martinez-Maldonado et al., 2017; Spikol et al., 2018; Ochoa et al., 2018 | Nakano et al., 2015 | Olsen et al. 2020; Spikol, Ruffaldi, Landolfi & Cukurova, 2017 | | 7 |
| Auditory + Visual + Physiological | Ouhaichi et al., 2021 | Neubauer et al., 2016; Peng & Nagao, 2021 | Reilly & Schneider, 2019 | | 4 |
| Auditory + Kinesthetic + Physiological | | | | Worsley & Blikstein, 2014; Madan et al., 2004 | 2 |
| Visual + Kinesthetic + Physiological | | | | Huang et al., 2019 | 1 |
| Visual + Tactile + Physiological | Fernandez-Nieto et al., 2021; Fernandez Nieto et al., 2022 | | | | 2 |
| Auditory + Visual + Kinesthetic + Physiological | | Echeverria et al., 2019 | Buckingham Shum et al., 2019 | | 2 |
| Visual + Kinesthetic + Physiological + Tactile | Martinez-Maldonado et al., 2020 | | | | 1 |
| **Total** | 11 | 11 | 9 | 5 | 36 |

Across all combinations, the MMCA studies used 12 different types of sensors. The most widely used — the microphone, which represents the auditory modality — featured in 27 studies (75%). Following closely were three sensors that represented the visual modality. Regular cameras were used in 17 studies (47%), specialized Kinect cameras with motion-tracking

capabilities were used in eight studies (22%), and cameras that track eye gaze were used in six studies (17%). In the tactile modality, two studies (5%) utilized computer keyboards that register pressure when students produce written output or perform actions within the software by pressing keys. Thirteen studies (36%) used tangible user interfaces, including tabletops or manikins, that register pressure when students perform certain actions (e.g., heart massage on a manikin, touching a specified area of a tabletop). Two studies (5%) utilized digital pens that register pressure when students produce written output. In the kinesthetic modality, five studies (14%) used accelerometers that allow detecting the movement of the whole body or body parts (e.g., wrist movements). Six studies (17%) used GPS trackers that enable the tracing of learner coordinates in a room. In the physiological modality, electrodermal activity (EDA) sensors, capturing physiological responses related to emotional arousal and stress, were utilized in 12 studies (33%). Other physiological sensors used were heartrate sensors (11%) and breathing rate sensors (3%). See Appendix 1 for a complete overview of the sensors.

The modalities were used to extract various types of information regarding the learners. Twenty-nine studies extracted data on participant speech by using the auditory modality, focusing on semantic content (N=12) and/or prosodic information (N=21). Twenty-one studies captured learner actions. This was done either through the visual modality by letting the computer or human annotator *see* which actions are performed (N=11) and/or through the tactile modality by letting the technology *pressure-sense* which actions are performed when learners press keyboards or tangible user interfaces (N=12). In addition, three studies used tactile sensors to extract information on the students' written output. Information on learner positions and movements was extracted in 21 studies. This involved either visual information from Kinect cameras (N=13) or kinesthetic information from accelerometers or GPS trackers (N=8). Furthermore, 11 studies collected information on eye gaze using the visual modality and six studied facial expressions using visual information. Finally, 14 studies drew on the physiological modality to extract either information on heartrate (N=4), skin conductance (N=12), and/or respiration rate (N=1). Figure 6 illustrates the complex relationship between modalities, sensors, and which information was extracted about the learner.

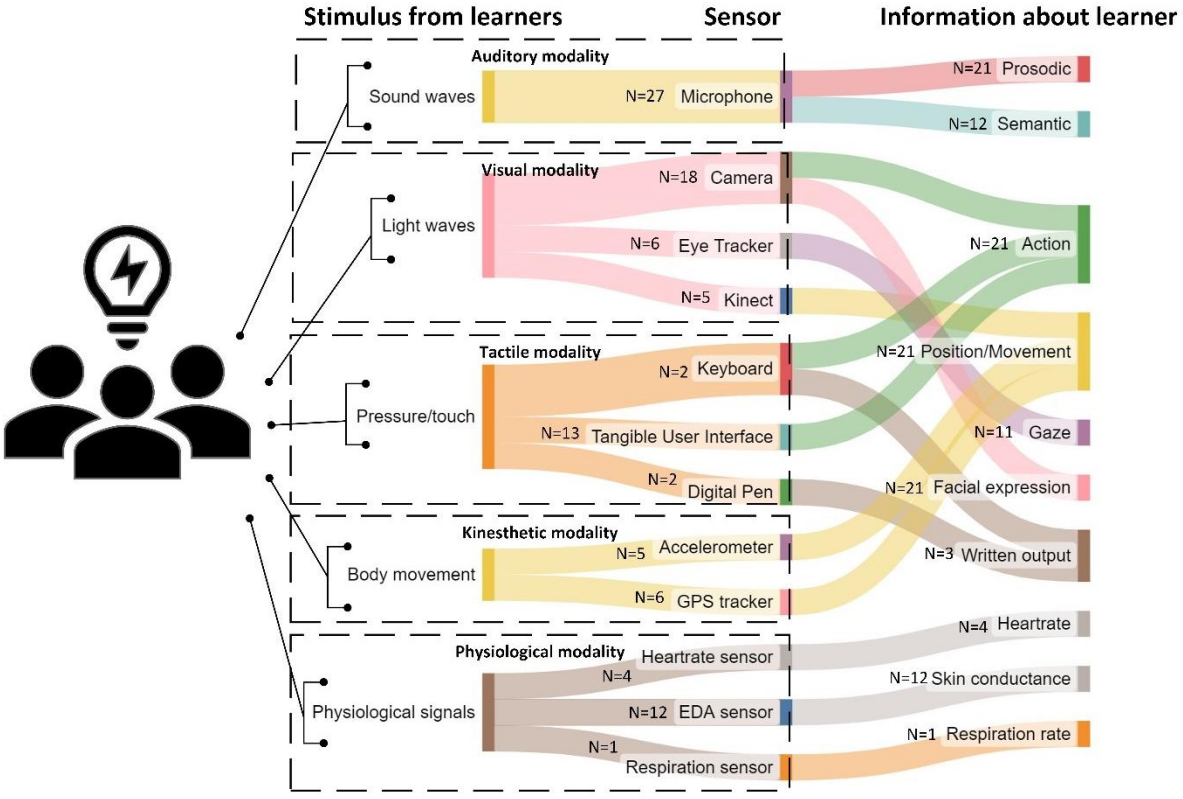


**Figure 6.** Relationship between modalities, sensors, and extracted information.

### 3.4. Rationales Underlying Combinations of Modalities in MMCA

After providing an overview of modalities and sensors, we now address our third research question. The review reveals that there are four main rationales for combining modalities in researching collaboration. Studies either take technical advancements as a point of departure (N=11), begin from a pedagogical/practical challenge (N=11), adopt a methodological focus (N=9), or begin from theoretical premises (N=5). For each rationale, we highlight a few of the empirical insights that MMCA studies and different combinations of modalities have added to our understanding of collaboration (see Table 2 for an overview).

### 3.5. Combining Modalities Due to Technical Advancements

Maybe not surprisingly considering the extensive use of technology in MMCA research, 11 studies can be described as being driven by technical innovations that enable new combinations of modalities. Taking Martinez-Maldonado et al. (2013) as an example, this study explores the use of tabletops with touchscreens to combine three modalities to further understand and enhance collaborative learning processes. The tabletops can automatically and unobtrusively collect auditory information on speech prosodics, visual information on participant movements, and tactile information from participant touches of the screen during a collaborative activity. Taking advantage of the synchronized multimodal data streams, the authors apply a sequential data mining approach to identify patterns of verbal participation and meaningful actions performed during group activity. When comparing these patterns against the actual quality of collaboration, as rated by two human raters, it was possible to differentiate between low and high collaborative groups just based on the data generated by the tabletops. Further, the technical potentials of tangible user interfaces were explored by combining tactile modalities with auditory and visual modalities. For example, Ochoa et al. (2018) showcase how tactile actions on a tabletop combined with semantic conversation data as well as visual data on facial expressions and gaze direction could be translated into multimodal transcripts. The authors illustrate how these transcripts are capable of informing teachers regarding the group's rapport and the quality of their collaborative efforts.

Another category of technology-driven research is studies that develop new tools with the aim of researching or facilitating collaboration. One of the earliest attempts is that by Sturm et al. (2007), who developed a prototype that collects and provides feedback on speaking data and visual information on participant head movements as a proxy for their eye gaze during group meetings. The study finds that real-time feedback on speaking time affects the group's social dynamics, while feedback on head orientation does not affect the group members' visual attention given to each other, as expected. Another example is the tool developed as part of the PELARS project (Spikol, Ruffaldi & Cukurova, 2017; Spikol et al., 2018). This customized MMCA tool can visually track movements of hands, faces, and objects in addition to using data from Arduino hardware that can turn tactile input into digital signals. In their studies, groups of engineering students were working on open-ended projects with the goal of creating physical artefacts. By applying machine learning (ML) to the multimodal data streams from the group work, the authors succeed in predicting the human rater's assessments of the quality of the group's artefacts.

A third category of technology-driven studies are those focused on further advancing existing technology by improved data visualizations. For example, Fernandez-Nieto et al. (2021, 2022) recognized the challenges associated with the interpretation of MMCA information by teachers and suggest the use of different visualizations to communicate insights to students. Based on foundations in data storytelling, the authors develop three interfaces (visual data slices, a tabular visualization, and written report) for visualizing logs and physiological data from co-located collaborative learning classes. The multimodal representations are found to be effective in supporting teacher reflection. Similarly, Martinez-Maldonado et al. (2017, 2020) explored the potential for using manikins and multimodal sensors in healthcare simulations to capture various facets of the collaborative learning process and to use personalized data storytelling approaches based on the visualization of learning analytics data for decision-making in the educational sector.

Overall, these articles demonstrate how technical premises often serve as a foundation for studies in MMCA that aim to develop technology and user interfaces that can advance our understanding of collaborative learning processes. In particular, the information from tactile interactions — whether it be through tabletops or manikins — into the multimodal analysis has proven instrumental in providing nuanced insights into the group dynamics, coherence, and outcomes of collaborative activities. These investigations demonstrate a clear trajectory towards more sophisticated, context-rich analyses of collaborative learning that leverage the concurrent application of multiple modalities for a deeper understanding and support of collaboration.

### 3.6. Combining Modalities to Address Pedagogical/Practical Challenges

Apart from studies with a technical rationale, our sample includes 11 studies that outline a practical challenge in educational and collaborative contexts as their main rationale for adopting a multimodal approach. One of the typical practical challenges mentioned in these studies is related to teachers having limited insight into collaborative learning, particularly when faced with large groups or online learning. Vrzakova et al. (2020) begin by describing the practical challenge of educators following up large classes that work remotely on collaborative problem-solving (CPS) tasks. In response, they strive to develop methods of combining different automatically observable "unimodal primitives" (e.g., student speech, body movement) to predict both subjective and objective outcomes of group CPS. Following this rationale, the authors combine visual screen recordings of online meetings and semantic audio data during the group work to gain insight into whether participants were talking or performed certain actions on their shared screen. The authors find that group silence correlates negatively with performance when only audio data is used. However, when utilizing visual data, silence combined with activity on the shared screen shows a positive correlation with performance. These results reveal how multimodality can help to differentiate our understanding of the problem when addressing practical challenges related to collaborative learning.

The practical challenge of limited resources for group work guidance is also addressed by Peng and Nagao (2021), who outline the difficulty of teachers monitoring the complex mental states of students working in groups. The authors develop a

multimodal tool that collects physiological heartrate data, visual data on facial expressions, hand movements, and prosodic speech data through sensors of commercial products (such as iPhone, AirPods, Apple Watch). The tool is then trained with supervised ML models to recognize the students' mental states, and the results are compared against external ratings of the mental states by two annotators. The authors find that combining physiological and auditory modalities yields better results for identifying states of confusion and concentration, while fusing all three modalities yields better results for detecting frustration and boredom. These results illustrate that combining more modalities does not always lead to better understanding, but the added value needs to be weighed against feasibility and research aims.

Another practical challenge taken up in the literature is related to the conditions within the groups that may affect the quality of collaboration. Neubauer et al. (2016) outline the problem of groups working on cognitively demanding tasks, which is associated with task-related stress that can impair performance and have lasting effects on health and emotional well-being. As this form of stress is typically associated with team cohesion, the authors develop an intervention to help groups build familiarity among team members before the task. They combine data from audio signals, facial expressions, and heart rate variability to identify signs of team cohesion and test the effect of their intervention on emotional resilience during group response to stressful tasks. The study provides an example of how combinations of modalities can be used to investigate practical questions related to group work that would otherwise be difficult to address.

In a similar vein, Müller et al. (2018) describe the practical challenge of building rapport in groups defined as "the close and harmonious relationship in which interaction partners are 'in sync' and can interact naturally and smoothly with each other" (p. 153). Since low rapport is often related to interpersonal conflicts during group work, it is relevant for teachers to be able to detect it and intervene to help groups avoid conflicts. To address this problem, the authors propose a multimodal approach for low rapport detection based on audio and video data comprising facial expressions and orientations, hand motion, speech activities, and prosodic features. They find that certain facial features related to negative emotions (e.g., nose wrinkling, brow-raising) are also associated with low rapport but that additional information from the other modalities did not further improve predictive power in this regard. This illustrates how important it is to critically assess which sets of features are relevant to extract from different modalities depending on the context.

Across these studies, a consistent theme is the value added by multimodal data in improving and facilitating collaborative processes in practice. The indicators obtained from various modalities serve not only as predictors of group dynamics and outcomes but are also useful tools for real-time intervention to guide and improve collaborative learning. Whether the focus is on identifying productive behaviours, mitigating dominance, or predicting potential group conflicts, multimodal synthesis provides a nuanced perspective that can foster more effective interactions among learners.

### 3.7. Combining Modalities Driven by Methodological Innovation

Another rationale for studying collaboration through different modalities is of a methodological nature. In our review, nine studies seek to leverage methodological innovations to enhance understanding of the collaborative learning process.

One aspect of these studies emerges from a specific interest in linguistic analyses, which are inherently linked to the auditory modality. For example, Schneider and Pea (2015) highlight the potential of natural language processing (NLP) in analyzing student discourse and provide a comprehensive methodological study that illustrates how NLP methods, in combination with joint visual gaze analyses, can predict "productive collaborative learning markers" during problem-solving. Empirically, the study investigates whether seeing one another's gaze affects verbal interaction. The authors find that the linguistic coordination and coherence by which students build on each other's ideas is indicative of productive collaborative learning, as rated by an external observer. The predictive power of linguistic coherence is further strengthened by including data on joint visual attention.

In addition to studies combining linguistic analysis with methods of analyzing visual data, other authors explore combinations with physiological or kinesthetic data. Reilly and Schneider (2019) explore how the NLP method Coh-Metrix can identify linguistic features of student speech during collaborative problem-solving (CPS) and which of the resulting metrics are most meaningful for estimating a group's collaboration. For this purpose, they collected additional electrodermal and eye gaze information to identify measures that have previously been associated with collaboration quality, such as joint visual attention or physiological synchrony. By comparing these measures with the results of their linguistic analysis, the authors can predict moments of productive collaboration among participants working on a programming task.

Driven by the possibilities of combining audio and spatial data modelling, Zhao et al. (2022) utilize auditory and kinesthetic data from nursing simulations to examine whether they could identify non-verbal events that can serve as meaningful indicators of student group performances and that are aligned with formal learning outcomes. The results reveal that analyzing speech events in relation to spatial position data indicates student role comprehension while they move through the room. In a subsequent study, Zhao et al. (2023) explore the potential of epistemic network analysis (ENA) to study kinesthetic and audio data from collocated teamwork in nursing simulations. They use ENA to identify key differences in communication behaviour related to the group's embodied teamwork performance. In the same vein, Buckingham-Shum et al. (2019) adopt a

methodological approach by operationalizing principles of "quantitative ethnography" (QE) and ENA to illustrate how multimodal data from nursing simulations may be analyzed.

All these studies are either inspired by applying current methods, such as ENA and NLP, or have the explicit ambition of further advancing analytical methods utilized in MMCA. Considering how speech is a rich source of information regarding collaboration, it is not surprising that a substantial amount of focus is placed on approaches to automatically analyze auditory data. However, there are also increasing attempts to supplement these speech analyses with other modalities. This development is related to the fact that it is increasingly accepted that focusing merely on verbal behaviour is not sufficient to understand collaboration contexts that require coordinated action.

### 3.8. Combining Modalities Guided by Theoretical Interest

While none of the reviewed studies have a purely theoretical focus, we identified five that can be characterized as guided by a theoretical interest in collaboration. Typically, these studies begin with a phenomenon or concept with origins in the wider literature on collaboration. Often, these concepts are then used to make sense of an existing multimodal data set or, vice versa, and the authors explore how the combination of modalities can enrich understanding of the concept.

One example is Ma et al.'s study (2022) that builds on a theoretical interest in the phenomenon of an impasse, which occurs when group members have an insufficient number of ideas or differing opinions that hinder them from progressing in their task. This study addresses the question of whether impasse during CPS can be automatically detected by ML applied to a data set comprising semantic and prosodic data as well as visual data on gaze, movement, and facial expression. Based on transcripts of the audio data, the authors develop a method to annotate turn exchanges that indicate that students have different opinions or insufficient ideas to progress, thereby indicating moments of impasse. Using the manually annotated transcript as a test data set, the ML model accurately classifies impasse moments and a large part of the variation is explained by the synchrony of the voice pitch and facial expressions of learners, which reveals the added value of combining auditory and visual modalities. In turn, the results contribute to enriching the theoretical understanding of an impasse as an empirical phenomenon that is not only detectable through verbal but also non-verbal indicators.

Another example of a study adopting a theoretical approach is that of Sinclair and Schneider (2021), who ground their study in the concept of *joint problem spaces* (Roschelle & Teasley, 1995). Their research examines how these spaces are constructed through alignment across various modes of communication. The authors develop "theoretically motivated metrics" (p. 431) to capture synchronization and alignment of linguistic and movement patterns of dyads working on CPS tasks. The extent of collaboration is assessed by two raters watching the dyads, while the MMCA data includes audio and visual data on hand movements. Not only does the study reveal that the level of linguistic and gestural alignment among learners is positively correlated to collaboration, it also finds that a combination of modalities yields better predictions than using data from each modality individually. The findings indicate that the creation of joint problem spaces not only involves alignment on a verbal plane but also on a non-verbal plane.

While these examples reveal a few studies that adopt a more theoretical starting point in our sample, other rationales dominate the MMCA literature. In particular, in the light of the theoretically well-founded educational research tradition on collaborative learning, it is surprising that not many conceptual foundations are used for designing MMCA systems or for providing more theoretical arguments for combining certain modalities to study collaboration in different contexts.

## 4. Discussion

In this scoping review, we provided a descriptive overview of MMCA studies and developed a definitional framework for the different modalities used in the literature. Based on the framework, we synthesized the findings of studies that employ different combinations of modalities and identified the underlying premises for conducting MMCA research. While the field is diverse and developing at a fast pace, engaging researchers with different backgrounds and varying interests in collaboration, a few general trends are evident, and these are explained below.

### 4.1. Definition and Categorization of Modalities

The first research question targeted the way modalities used in MMCA studies can be defined and categorized. Our review illustrates that MMCA research is not driven by a single perspective but embraces a spectrum of motivations, ultimately enriching the field with varied insights and approaches. Meanwhile, we argue that the diversity of premises in MMCA research challenges the establishment of standardized methodologies. The tendency to utilize different terminology and recreating measures that have already been developed elsewhere generate issues when generalizing results across literature. Not least, this diversity blurs the basis for comparability across studies and hinders the development of cohesive theories in the field, as an integrative understanding of empirical findings is relatively limited. In line with Schneider et al. (2021), we advocate for unifying the naming of modalities and their measures across studies.

To this end, our study proposed a framework comprising five main categories (i.e., visual, auditory, tactile, kinesthetic, and physiological) for defining and categorizing the most common modalities used in MMCA. The framework contributes to

a more integrated understanding of the common endeavour of MMCA by providing a common terminology and rationale. The results make it easier for future MMCA researchers to distinguish the types of modalities that are combined and to explicate the rationales for selecting these combinations. Standardized definitions of modality, stimulus, and sensors lay the groundwork for clearer communication and more consistent reporting in MMCA research as well as clear the space for the potential future use of other modalities. This, in turn, facilitates collaboration and knowledge-sharing among researchers in MMCA.

By providing our definitional framework, we acknowledge the complexity of the concept of modality and explicitly position our definition against other possible ways of defining the term. Defining modality as a relationship between stimulus and sensor requires us to consider who is the stimulus and who is the sensor. As our framework is aimed at MMCA research, we focus on the *relationship between the groups of participants as the objects of study (i.e., those who generate the stimulus), and the researcher who captures the stimuli generated by these participants (i.e., by using their own senses or technical sensors)*. Consequently, our choice of terminology is derived from this perspective; for example, by referring to a keyboard "sensing the participants' touch" as a tactile modality. Alternatively, and maybe more common in everyday language, we can think of modality as referring to *the relationship between stimuli in the external world and how the participants themselves sense and engage with these stimuli*. Staying with the example of tactile modality, this perspective draws our focus on the participants being the ones sensing touch from external stimuli, such as fellow students or objects they use. While this is a legitimate perspective, particularly in research that attempts to describe the way in which participants perceive and interact with each other and their learning environment, we argue that it is more important for MMCA researchers to understand their own and their tools' modal relationships with the objects of their study.

By highlighting the differences among different sensory modalities, our framework sets the stage for a more nuanced analysis of multimodal data. Researchers can now systematically explore the unique aspects of each modality, thereby leading to a deeper understanding of how to capture group behaviour and interactions during collaboration. For example, in auditory modality, the framework highlights the potential for distinguishing prosodic and semantic aspects, providing insights into both "how" and "what" students communicate. Similarly, the framework makes it possible to deliberate which modality to choose when aiming to extract a specific type of information from the group. For example, when researchers aim to capture participant movements, they can consider data from the visual modality using Kinect cameras or data from the kinesthetic modality using accelerometers attached to the students' bodies. Considering the complexity of the field, we argue that only a systematic and transparent documentation of sensors and modalities makes it possible to advance the understanding of which combinations are meaningful to study and support collaboration and acknowledging cases in which data from additional modalities is possibly not useful and, therefore, should not be further pursued.

### 4.2. Combinations of Modalities and Sensors

The second research question sought to explore the combinations of modalities and sensors used in MMCA studies. The analysis revealed a diversity in modalities and sensor combinations, which can be attributed to the ongoing evolution of the field, characterized by a quest for the most effective approaches to understanding collaborative learning. This diversity could indicate that MMCA research is still in the process of maturation and lacks a consolidated body of knowledge regarding which modality combinations are most effective. In addition, the complex and dynamic nature of the collaborative processes often necessitates a multidimensional perspective, one that can uncover hidden patterns and interactions that remain concealed when examining individual modalities in isolation. These intricate processes may involve emotional states, cognitive dynamics, and social interactions, all of which can be more effectively captured through multimodal analysis. This sophistication is exemplified in the work of Vrzakova et al. (2020), which showcases the power of combining various modalities to obtain a more comprehensive understanding of learner behaviours and cognitive processes during collaboration.

The multidisciplinary nature of MMCA further complicates the matter. Different researchers — from diverse academic backgrounds such as computer science (e.g., Spikol et al., 2018; Martinez-Maldonado et al., 2017), learning sciences, or educational psychology (e.g. Worsley & Blikstein, 2014; Malmberg et al., 2019) — may rely on distinct sets of modalities to comprehend collaborative processes. This multidisciplinarity implies MMCA's adaptability and its capacity to cater to the unique requirements and perspectives of various researchers. Consequently, while the current landscape of MMCA research shows progress and innovation in combining modalities, it also hints at the need for future studies to delve deeper into the effectiveness and applicability of different combinations.

The application of multiple modalities to MMCA offers the potential for advanced modelling and enhanced insights into collaborative learning. By accessing a broader and diversified data set, researchers can move beyond simple, isolated observations and delve into the subtleties of how collaboration unfolds. As argued by Reimann et al. (2014), multimodal data provides access to hard-to capture or even invisible responses by the learners' body and brain, thereby enriching the ontologically flat data available through self-reports or mere observations. This multifaceted approach aligns with the holistic nature of collaborative learning and reflects the dynamic character of the field. The rich variety of modalities utilized in MMCA research is indicative of the field's potential to offer a more profound understanding of the complex and multifaceted nature of collaborative learning. As MMCA research continues to develop, it is likely that further advancements in data integration

and analysis will provide even more sophisticated models and insights, ultimately benefiting both researchers and educators in their quest to understand and improve collaborative learning experiences.

Similar to previous reviews (Praharaj et al., 2021), we found a widespread utilization of audio and visual modalities, underscoring the significance of these sensory inputs in the realm of MMCA. The study conducted by Sturm et al. in 2007 serves as a pioneering example, illustrating the early recognition of auditory and visual modalities as potential tools for investigating social dynamics in collaborative environments. Their work, which provided real-time feedback based on speaking time and speaker dominance, underscores how auditory and visual modalities can effectively capture essential cues related to the quality of collaboration. It is therefore no surprise that more than half of the included studies used this combination, either alone or together with additional modalities.

Furthermore, the accessibility and cost-effectiveness of auditory and visual modalities may provide a compelling rationale for their frequent use in MMCA. Researchers can readily harness these modalities to analyze text and video actions, thereby facilitating the extraction of valuable insights pertaining to collaboration. In contrast, physiological modalities — such as heart rate and skin conductance — often necessitate expensive sensors, and the data they generate can be challenging to collect, analyze, and relate to learning constructs. In Europe, the use of health data is also highly regulated by data privacy laws. These practical considerations drive researchers towards employing audio and visual modalities, which provide a more economical and interpretable means of studying collaboration dynamics. Nonetheless, the review reveals a considerable use of wearable devices, such as EDA and heart rate trackers, which could be attributed to the increasing need to capture and analyze non-verbal signals and actions during collaborative tasks (e.g., stress; Noel et al., 2022) and how these might affect the collaboration process.

### 4.3. Empirical Contributions and Underlying Rationales

Addressing the third research question, we found that the empirical contributions and underlying rationales reflect a diverse landscape. A common foundation for the studies is often grounded in practical–pedagogical or technical premises, with the purpose of developing technology and user interfaces that can support collaborative learning processes in practice. As research in this domain is frequently driven by novel technical or methodological innovations, MMCA researchers strive to explore their potential for advancing collaboration. The diverse premises demonstrate the adaptability of MMCA and its ability to integrate technology, methodology, and practical pedagogy. In terms of the theoretical characteristics of MMCA studies, like previous studies (e.g., Giannakos & Cukurova, 2023), we found a limited utilization of theory in most of the analyzed MMCA studies. Given that collaboration is the main focus of study, a range of theoretical approaches and concepts from the field of education, psychology, and sociology could be relevant for MMCA research. Nevertheless, the majority of the reviewed studies exhibit a technical, pedagogical, or methodological focus, primarily aiming to develop, assess, or refine tools, techniques, and approaches pertinent to MMCA. The prioritization of the practical application of MMCA over theoretical abstraction is arguably due to the interdisciplinary nature of MMCA research, which draws upon a diverse range of disciplines, including computer science and sensor technology. For example, studies that are deeply rooted in a technical disciplinary context may emphasize practical tools and applications, while others may leverage established theories from social science disciplines to inform their research. These interdisciplinary dynamics emphasize the multifaceted nature of MMCA research, with theoretical grounding often contingent on the specific aims and disciplinary perspectives of individual studies.

Another explanation for the limited attention to theory is related to the fact that MMCA research is characterized by a wide use of ML methods. Most examples of studies using ML are supervised approaches where an algorithm learns to identify patterns that can distinguish high from low collaboration quality (e.g., Spikol et al., 2018). For such approaches, it is necessary to have a certain pre-defined outcome variable that typically includes some form of assessment (e.g., a human annotator assessing collaboration quality, pre–post learning test results). Even if the machine succeeds in categorizing groups correctly based on the multimodal data, the algorithm typically remains a black box that does not provide any insight into what exactly the machine identified as relevant in the data set to discern high- from low-performing groups. From a theoretical perspective, the results from such approaches are problematic, as they provide little added value to our conceptual understanding of how and why collaboration succeeds.

As expected, most studies primarily concentrate on understanding or advancing collaborative work or learning. Nonetheless, we observe that MMCA is also exploited to address distinctly different research questions. For example, Neubauer et al. (2016) investigate the implications of group coherence, measured by MMCA methods, on student stress and resilience levels while performing challenging tasks. This can be an indication that the methodologies developed in MMCA research have the potential to enrich the empirical study of a diversity of psychological and educational phenomena that are also relevant beyond the focus on collaborative learning.

In general, our review reflects an optimistic view of the future of MMCA, with most contributions indicating the promising prospects of using multiple modalities. In certain instances, clear arguments for selected modalities are elucidated and the necessary inclusion of several modalities well argued for (e.g., Malmberg et al., 2019). However, others lack explicit rationales for selecting specific modalities. In such cases, choices might be attributed to the availability of certain technologies or

competences in a research group. With MMCA still being in an explorative stage, this is not overly surprising. However, the field's advancement could be accelerated by more explicit reporting on the underlying reasons for pairing modalities and a more explicit anchoring in established conceptualizations of collaboration.

Within the review, there is also a subset of studies that do not provide evidence of the enhanced value of using multiple modalities (e.g., Viswanathan & VanLehn, 2018). Considering the potential intrusion on participant privacy and the need for immense data processing power, it is particularly important to consider the need for multimodal data. However, under closer scrutiny, it becomes evident that the ultimate determination of the added value depends strongly on the underlying rationales of the studies. For example, in studies attempting to differentiate high versus low collaboration using multimodal data sets, only certain modalities may demonstrate relevant predictive power (e.g., Martinez-Maldonado et al., 2013; Viswanathan & VanLehn, 2018). Conversely, studies that aim to achieve a nuanced understanding of collaboration processes might find the same combination of modalities valuable.

Finally, our review also finds notable technical limitations within MMCA studies and certain types of data not (yet) being amenable to automatic processing. As of today, these challenges are often resolved by a human interpretation intermediary, either by logging observed actions in real-time or retrospective transcript annotation (e.g., Fernandez Nieto et al., 2022). In these cases, it could be argued that humans act as "sensors" within the MMCA system by lending their eyes or ears to the otherwise technical set of sensory devices. Compared to technology, human sensors are highly versatile and can extract several types of data on learners simultaneously; for example, if the same human logs student gaze direction (modality: visual) and whether a student is speaking (modality: auditory). Human annotators still play an important role in much of current MMCA literature, but there is a general ambition to automatize both data collection and data analysis to reduce the need for human involvement to the bare minimum. While we stand in accord with the immense potential of MMCA to enhance our theoretical and practical comprehension of collaborative learning, we also advocate for a measured approach. It is our assertion that a more critical discourse, addressing the ethical and environmental ramifications of fully automating MMCA, is of paramount importance.

### 4.4. Limitations

While this scoping review provides valuable insights into the domain of MMCA, there are certain limitations that should be acknowledged: First, the study employed a strict cut-off date of January 2023. This implies that more recent developments and studies in MMCA beyond that date are not included in this review. Second, while the review encompassed 36 studies, this sample size — though indicative of trends — may not fully capture the breadth of MMCA research. The field is rapidly evolving, and future research may reveal additional insights. Third, the interdisciplinary nature of MMCA can lead to varied definitions, methodologies, and objectives across studies. While we have attempted to categorize modalities and provide clarity, the field's diverse nature may still pose challenges in achieving a wholly unified framework. Moreover, the categorization of studies based on theoretical, technical, methodological, or pedagogical rationales was based on our interpretation of the studies' starting points and rationales. These were, for the most part, stated explicitly, while a few studies might have more implicit premises, which could affect the comprehensive understanding of the rationales guiding MMCA research. In addition, the use of ACM and IEEE databases may have led to an overrepresentation of studies taking a more technical focus in our sample. Despite these limitations, this scoping review serves as a foundational exploration of MMCA, offering a platform for researchers, technology developers, and practitioners to build upon and further refine our understanding of this evolving field.

## 5. Conclusion

In the ever-evolving domain of MMCA, this comprehensive review has not only unveiled the state of this field but has also sought to contribute to a deeper understanding of its nature and core challenges. Foremost among these contributions is the development of a comprehensive framework for defining and categorizing modalities. This framework responds to the pressing need for clarity and an integrated approach in a field characterized by the diverse and multifaceted use of modalities in MMCA research. By categorizing modalities into five general categories — auditory, visual, physiological, kinesthetic, and tactile — this framework serves as a foundational guide for researchers, designers, and practitioners, thereby fostering a unified understanding of MMCA. Furthermore, our investigation has mapped the geographical landscape of MMCA research, pinpointing the countries where it is most active, with a prominent presence in the US and Australia, along with clusters in various European countries. Understanding this geographical distribution is important for collaboration, knowledge sharing, and the internationalization of MMCA studies. Additionally, we have identified the diverse combinations of modalities prevalent in MMCA research, ranging from two to four modalities. The prominence of audio and visual modalities in these combinations emphasizes their significance in capturing the complexities of collaborative learning, but also indicates existing limitations in technology. This knowledge enables researchers and practitioners to make informed decisions regarding the modalities that best suit their specific research objectives. Lastly, our exploration into the theoretical, technical, methodological, and pedagogical premises guiding MMCA studies provides vital insights into the rationales and approaches

within the field. Recognizing this diversity of rationales is essential for understanding the multifaceted nature of MMCA research and fostering collaboration among stakeholders. These findings are important to researchers, MMCA tool designers, and practitioners, as they equip them with the knowledge required to navigate the complexities and potentials of collaborative learning in a multimodal world. The insights generated by this study have the potential to pave the way for more effective MMCA interventions and systems, thereby contributing to the continued growth and impact of the field.

## Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The publication of this article received financial support from the Norwegian Research Council (grant number: 324826).

## References

Andriessen, J., & Baker, M. (2020). *On collaboration: Personal, educational and societal arenas*. Sense-Brill.

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology, 8*(1), 19–32. https://doi.org/10.1080/1364557032000119616

Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics, 3*(2), 220–238. https://doi.org/10.18608/jla.2016.32.11

Booth, A., Sutton, A., Clowes, M., & Martyn-St James, M. (2022). *Systematic approaches to a successful literature review* (3rd ed.). SAGE Publications.

*Buckingham Shum, S., Echeverria, V., & Martinez-Maldonado, R. (2019). The multimodal matrix as a quantitative ethnography methodology. In B. Eagan, M. Misfeldt, & A. Siebert-Evenstone (Eds.), *Advances in quantitative ethnography: First international conference, ICQE 2019, Madison, WI, USA, October 20–22, 2019, proceedings* (pp. 26–40). Springer Cham. https://doi.org/10.1007/978-3-030-33232-7_3

Bunt, H., Beun, R.-J., & Borghuis, T. (1998). *Multimodal human–computer communication: Systems, techniques, and experiments.* Springer Berlin, Heidelberg. https://doi.org/10.1007/BFb0052309

*Chejara, P., Prieto, L. P., Ruiz-Calleja, A., Rodríguez-Triana, M. J., Shankar, S. K., & Kasepalu, R. (2020). Quantifying collaboration quality in face-to-face classroom settings using MMLA. In A. Nolte, C. Alvarez, R. Hishiyama, I.-A. Chounta, M. J. Rodríguez-Triana, & T. Inoue (Eds.), *Collaboration technologies and social computing: 26th international conference, CollabTech 2020, Tartu, Estonia, September 8–11, 2020, proceedings* (pp. 159–166). Springer Cham. https://doi.org/10.1007/978-3-030-58157-2_11

Chua, Y. H. V., Dauwels, J., & Tan, S. C. (2019). Technologies for automated analysis of co-located, real-life, physical learning spaces: Where are we now? In C. Brooks, R. Ferguson, & U. Hoppe (Program chairs), *Learning Analytics to Promote Inclusion and Success: Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 11–20). ACM Press. https://doi.org/10.1145/3303772.3303811

*Echeverria, V., Martinez-Maldonado, R., & Buckingham Shum, S. (2019). Towards collaboration translucence: Giving meaning to multimodal group data. In S. Brewster, G. Fitzpatrick, A. Cox, & V. Kostakos (Chairs), *CHI 2019: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper 39). ACM Press. https://doi.org/10.1145/3290605.3300269

*Fernandez Nieto, G. M., Kitto, K., Buckingham Shum, S., & Martinez-Maldonado, R. (2022). Beyond the learning analytics dashboard: Alternative ways to communicate student data insights combining visualisation, narrative and storytelling. In A. F. Wise, R. Martinez-Maldonado, & I. Hilliger (Program chairs), *Learning Analytics for Transition, Disruption and Social Change: The Twelfth International Conference on Learning Analytics & Knowledge* (pp. 219–229). ACM Press. https://doi.org/10.1145/3506860.3506895

*Fernandez-Nieto, G. M., Echeverria, V., Buckingham Shum, S., Mangaroska, K., Kitto, K., Palominos, E., Axisa, C., & Martinez-Maldonado, R. (2021). Storytelling with learner data: Guiding student reflection on multimodal team data. *IEEE Transactions on Learning Technologies, 14*(5), 695–708. https://doi.org/10.1109/TLT.2021.3131842

*Huang, K., Bryant, T., & Schneider, B. (2019). Identifying collaborative learning states using unsupervised machine learning on eye-tracking, physiological and motion sensor data. In C. F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.), *Proceedings of the 12th International Conference on Educational Data Mining* (pp. 318–323).

Giannakos, M., & Cukurova, M. (2023). The role of learning theory in multimodal learning analytics. *British Journal of Educational Technology, 54*(5), 1246–1267. https://doi.org/10.1111/bjet.13320

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

Fritzsch, B. (Ed.). (2021). *The senses: A comprehensive reference* (2nd Ed.). Elsevier.

Graesser, A. C., Greiff, S., Stadler, M., & Shubeck, K. T. (2020). Collaboration in the 21st century: The theory, assessment, and teaching of collaborative problem solving. *Computers in Human Behavior, 104*, Article 106134. https://doi.org/10.1016/j.chb.2019.09.010

Kim, Y.-J., & Yoo, J.-H. (2020). The utilization of debriefing for simulation in healthcare: A literature review. *Nurse Education in Practice, 43*, Article 102698. https://doi.org/10.1016/j.nepr.2020.102698

*Kim, T., Chang, A., Holland, L., & Pentland, A. S. (2008). Meeting mediator: Enhancing group collaboration using sociometric feedback. In B. Begole & D. W. McDonald (General chairs), *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (pp. 457–466). ACM Press. https://doi.org/10.1145/1460563.1460636

*Ma, Y., Celepkolu, M., & Boyer, K. E. (2022). Detecting impasse during collaborative problem solving with multimodal learning analytics. In A. F. Wise, R. Martinez-Maldonado, & I. Hilliger (Program chairs), *Learning Analytics for Transition, Disruption and Social Change: The Twelfth International Learning Analytics & Knowledge Conference* (pp. 45–55). ACM Press. https://doi.org/10.1145/3506860.3506865

Mao, A., Mason, W., Suri, S., & Watts, D. J. (2016). An experimental study of team size and performance on a complex task. *PLoS ONE, 11*(4), Article e0153048. https://doi.org/10.1371/journal.pone.0153048

*Madan, A., Caneel, R., & Pentland, A. S. (2004). GroupMedia: Distributed multi-modal interfaces. *Proceedings of the 6th International Conference on Multimodal Interfaces* (pp. 309–316). ACM Press. https://doi.org/10.1145/1027933.1027983

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

*Martinez-Maldonado, R., Dimitriadis, Y., Martinez-Monés, A., Kay, J., & Yacef, K. (2013). Capturing and analyzing verbal and physical collaborative learning interactions at an enriched interactive tabletop. *International Journal of Computer-Supported Collaborative Learning, 8*(4), 455–485. https://doi.org/10.1007/s11412-013-9184-1

*Martinez-Maldonado, R., Echeverria, V., Fernandez Nieto, G., & Buckingham Shum, S. (2020). From data to insights: A layered storytelling approach for multimodal learning analytics. In R. Bernhaupt, F. Mueller, D. Verwrij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjørn, S. Zhao, B. P. Samson, & R. Kocielnik (Chairs), *CHI'20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Paper 21). ACM Press. https://doi.org/10.1145/3313831.3376148

Martinez-Maldonado, R., Gašević, D., Echeverria, V., Fernandez Nieto, G., Swiecki, Z., & Buckingham Shum, S. (2021). What do you mean by collaboration analytics? A conceptual model. *Journal of Learning Analytics, 8*(1), 126–153. https://doi.org/10.18608/jla.2021.7227

Martinez-Maldonado, R., Kay, J., Buckingham Shum, S., & Yacef, K. (2019). Collocated collaboration analytics: Principles and dilemmas for mining multimodal interaction data. *Human–Computer Interaction, 34*(1), 1–50. https://doi.org/10.1080/07370024.2017.1338956

*Martinez-Maldonado, R., Power, T., Hayes, C., Abdiprano, A., Vo, T., Axisa, C., & Buckingham Shum, S. (2017). Analytics meet patient manikins: Challenges in an authentic small-group healthcare simulation classroom. In A. F. Wise, P. H. Winne, G. Lynch, X. Ochoa, I. Molenaar, S. Dawson, & M. Hatala (Chairs), *Understanding, Informing and Improving Learning with Data: The Seventh International Learning Analytics & Knowledge Conference* (pp. 90–94). ACM Press. https://doi.org/10.1145/3027385.3027401

Marlow, S., Bisbey, T., Lacerenza, C., & Salas, E. (2018). Performance measures for health care teams: A review. *Small Group Research, 49*(3), 306–356. https://doi.org/10.1177/1046496417748196

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group*. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), Article e1000097. https://doi.org/10.1371/journal.pmed.1000097

Muukkonen, H., Lakkala, M., Lahti-Nuuttila, P., Ilomäki, L., Karlgren, K., & Toom, A. (2020). Assessing the development of collaborative knowledge work competence: Scales for higher education course contexts. *Scandinavian Journal of Educational Research, 64*(7), 1071–1089. https://doi.org/10.1080/00313831.2019.1647284

Munn, Z., Peters, M. J. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology, 18*, Article 143. https://doi.org/10.1186/s12874-018-0611-x

*Müller, P., Huang, M. X., & Bulling, A. (2018). Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In S. Berkovsky, Y. Hijikata, J. Rekimoto, M. Burnett, M. Billinghurst, & A. Quigley (Chairs), *IUI 2018: Proceedings of the 23rd International Conference on Intelligent User Interfaces* (pp. 153–164). ACM Press. https://doi.org/10.1145/3172944.3172969

*Nakano, Y. I., Nihonyanagi, S., Takase, Y., Hayashi, Y., & Okada, S. (2015). Predicting participation styles using co-occurrence patterns of nonverbal behaviours in collaborative learning. In Z. Zhang, P. Cohen, D. Bohus, R. Horaud, & H. Meng (Eds.), *ICMI'15: Proceedings of the 2015 ACM International Conference on Multimodal Interaction* (pp. 91–98). ACM Press. https://doi.org/10.1145/2818346.2820764

*Neubauer, C., Woolley, J., Khooshabeh, P., & Scherer, S. (2016). Getting to know you: A multimodal investigation of team behavior and resilience to stress. In Y. I. Nakano, E. André, T. Nishida, L.-P. Morency, C. Busso, & C. Pelachaud (Chairs), *ICMI'16: Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 193–200). ACM Press. https://doi.org/10.1145/2993148.2993195

*Noroozi, O., Alikhani, I., Järvelä, S., Kirschner, P. A., Juuso, I., & Seppänen, T. (2019). Multimodal data to design visual learning analytics for understanding regulation of learning. *Computers in Human Behavior, 100*, 298–304. https://doi.org/10.1016/j.chb.2018.12.019

Noël, R., Miranda, D., Cechinel, C., Riquelme, F., Primo, T. T., & Munoz, R. (2022). Visualizing collaboration in teamwork: A multimodal learning analytics platform for non-verbal communication. *Applied Sciences, 12*(15), Article 7499. https://doi.org/10.3390/app12157499

*Ochoa, X., Chiluiza, K., Granda, R., Falcones, G., Castells, J., & Guamán, B. (2018). Multimodal transcript of face-to-face group-work activity around interactive tabletops. In A. Pardo, K. Bartimote, G. Lynch, S. Buckingham Shum, R. Ferguson, A. Merceron, & X. Ochoa (Eds.), *Companion Proceedings of the 8th International Conference on Learning Analytics & Knowledge*. SoLAR.

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

*Ouhaichi, H., Spikol, D., & Vogel, B. (2021). MBOX: Designing a flexible IoT multimodal learning analytics system. In M. Chang, N.-S. Chen, D. G. Sampson, & A. Tlili (Eds.), *2021 International Conference on Advanced Learning Technologies (ICALT)* (pp. 122–126). IEEE. https://doi.org/10.1109/ICALT52272.2021.00044

Oviatt, S., & Cohen, A. (2014). Written activity, representations and fluency as predictors of domain expertise in mathematics. In A. A. Salah, J. Cohn, B. Schuller, O. Aran, L.-P. Morency, & P. R. Cohen (Chairs), *ICMI'14: Proceedings of the 2014 International Conference on Multimodal Interaction* (pp. 10–17). ACM Press. https://doi.org/10.1145/2663204.2663245

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

*Peng, S., & Nagao, K. (2021). Recognition of students' mental states in discussion based on multimodal data and its application to educational support. *IEEE Access, 9*, 18235–18250. https://doi.org/10.1109/ACCESS.2021.3054176

*Praharaj, S., Scheffel, M., Drachsler, H., & Specht, M. (2018). Multimodal analytics for real-time feedback in co-located collaboration. In V. Pammer-Schindler, M. Pérez-Sanagustín, H. Drachsler, R. Elferink, & M. Scheffel (Eds.), *Lifelong technology-enhanced learning: 13th European conference on technology enhanced learning, EC-TEL 2018, Leeds, UK, September 3–5, 2018, proceedings* (pp. 187–201). Springer Cham. https://doi.org/10.1007/978-3-319-98572-5_15

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K. E., & Ansari, R. (2002). Multimodal human discourse: Gesture and speech. *ACM Transactions on Computer–Human Interaction, 9*(3), 171–193. https://doi.org/10.1145/568513.568514

*Reilly, J. M., & Schneider, B. (2019). Predicting the quality of collaborative problem solving through linguistic analysis of discourse. In C. F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.), *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)* (pp. 149–157). International Educational Data Mining Society. https://educationaldatamining.org/edm2019/proceedings/

Reimann, P., Markauskaite, L., & Bannert, M. (2014). e-Research and learning theory: What do sequence and process mining methods contribute? *British Journal of Educational Technology, 45*(3), 528–540. https://doi.org/10.1111/bjet.12146

Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In C. O'Malley (Ed.), *Computer supported collaborative learning* (pp. 69–97). NATO ASI Series, vol 128. Springer. https://doi.org/10.1007/978-3-642-85098-1_5

*Scherer, S., Weibel, N., Morency, L.-P., & Oviatt, S. (2012). Multimodal prediction of expertise and leadership in learning groups. *MLA'12: Proceedings of the 1st International Workshop on Multimodal Learning Analytics* (Article 1). ACM Press. https://doi.org/10.1145/2389268.2389269

*Schneider, B., & Pea, R. (2015). Does seeing one another's gaze affect group dialogue? A computational approach. *Journal of Learning Analytics, 2*(2), 107–133. https://doi.org/10.18608/jla.2015.22.9

Schneider, B., Sung, G., Chng, E., & Yang, S. (2021). How can high-frequency sensors capture collaboration? A review of the empirical links between multimodal metrics and collaborative constructs. *Sensors, 21*(24), Article 8185. https://doi.org/10.3390/s21248185

*Sinclair, A. J., & Schneider, B. (2021). Linguistic and gestural coordination: Do learners converge in collaborative dialogue? In S. Hsiao, S. Sahebi, F. Bouchet, & J.-J. Vie (Eds.), *Proceedings of the 14th International Conference on Educational Data Mining* (pp. 431–438). International Educational Data Mining Society. https://files.eric.ed.gov/fulltext/ED615472.pdf

*Spikol, D., Ruffaldi, E., & Cukurova, M. (2017). Using multimodal learning analytics to identify aspects of collaboration in project-based learning. In B. K. Smith, M. Borge, E. Mercier, & K. Y. Lim (Eds.), *Making a difference: Prioritizing equity and access in CSCL, 12th International Conference on Computer Supported Collaborative Learning (CSCL) 2017* (pp. 263–270). International Society of the Learning Sciences. https://repository.isls.org/handle/1/240

*Spikol, D., Ruffaldi, E., Landolfi, L., & Cukurova, M. (2017). Estimation of success in collaborative learning based on multimodal learning analytics features. In M. Chang, N.-S. Chen, R. Huang, Kinshuk, D. G. Sampson, & R. Vasiu (Eds.), *The 17th IEEE International Conference on Advanced Learning Technologies (ICALT 2017)* (pp. 269–273). IEEE. https://doi.org/10.1109/ICALT.2017.122

*Spikol, D., Ruffaldi, E., Dabisias, G., & Cukurova, M. (2018). Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *Journal of Computer Assisted Learning, 34*(4), 366–377. https://doi.org/10.1111/jcal.12263

*Stewart, A. E. B., Keirn, Z., & D'Mello, S. K. (2021). Multimodal modeling of collaborative problem-solving facets in triads. *User Modelling and User-Adapted Interaction, 31*(4), 713–751. https://doi.org/10.1007/s11257-021-09290-y

*Sturm, J., Herwijnen, O. H., Eyck, A., & Terken, J. (2007). Influencing social dynamics in meetings through a peripheral display. In K. Mase, D. Massaro, K. Takeda, D. Roy, & A. Potamianos (Chairs), *ICMI'07: Proceedings of the Ninth International Conference on Multimodal Interfaces* (pp. 263–270). ACM Press. https://doi.org/10/b3fcjc

Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving, *Computers & Education, 143*, Article 103672. https://doi.org/10.1016/j.compedu.2019.103672

Theobald, E. J., Eddy, S. L., Grunspan, D. Z., Wiggins, B. L., & Crowe, A. J. (2017). Student perception of group dynamics predicts individual performance: Comfort and equity matter. *PLoS ONE, 12*(7), Article e0181336. https://doi.org/10.1371/journal.pone.0181336

Turk, M. (2014). Multimodal interaction: A review. *Pattern Recognition Letters, 36*, 189–195. https://doi.org/10.1016/j.patrec.2013.07.003

*Viswanathan, S. A., & VanLehn, K. (2018). Using the tablet gestures and speech of pairs of students to classify their collaboration. *IEEE Transactions on Learning Technologies, 11*(2), 230–242. https://doi.org/10/gds4fq

*Vrzakova, H., Amon, M. J., Stewart, A., Duran, N. D., & D'Mello, S. K. (2020). Focused or stuck together: Multimodal patterns reveal triads' performance in collaborative problem solving. In C. Rensing, H. Drachsler, V. Kovanović, N. Pinkwart, M. Scheffel, & K. Verbert (Chairs), *Celebrating 10 years of LAK: Shaping the Future of the Field: The Tenth International Conference on Learning Analytics & Knowledge* (pp. 295–304). ACM Press. https://doi.org/10.1145/3375462.3375467

*Worsley, M., & Blikstein, P. (2014). Deciphering the practices and affordances of different reasoning strategies through multimodal learning analytics. In X. Ochoa, M. Worsley, K. Chiluiza, & S. Luz (Chairs), *MLA'14: Proceedings of the 2014 ACM Multimodal Learning Analytics Workshop and Grand Challenge* (pp. 21–27). ACM Press. https://doi.org/10.1145/2666633.2666637

Wise, A. F., Knight, S., & Buckingham Shum, S. (2021). Collaborative learning analytics. In U. Cress, C. Rosé, A. F. Wise & J. Oshima (Eds.), *International handbook of computer-supported collaborative learning* (pp. 425–443). Springer Cham. https://doi.org/10.1007/978-3-030-65291-3_23

*Zhao, L., Swiecki, Z., Gašević, D., Yan, L., Dix, S., Jaggard, H., Wotherspoon, R., Osborne, A., Li, X., Alfredo, R., & Martinez-Maldonado, R. (2023). METS: Multimodal learning analytics of embodied teamwork learning. In I. Hilliger, H. Khosravi, B. Rienties, & S. Dawson (Chairs), *Towards Trustworthy Learning Analytics: The Thirteenth International Conference on Learning Analytics & Knowledge* (pp. 186–196). ACM Press. https://doi.org/10.1145/3576050.3576076

*Zhao, L., Yan, L., Gašević, D., Dix, S., Jaggard, H., Wotherspoon, R., Alfredo, R., Li, X., & Martinez-Maldonado, R. (2022). Modelling co-located team communication from voice detection and positioning data in healthcare simulation. In A. F. Wise, R. Martinez-Maldonado, & I. Hilliger (Chairs), *Learning Analytics for Transition, Disruption and Social Change: The Twelfth International Conference on Learning Analytics & Knowledge* (pp. 370–380). ACM Press. https://doi.org/10.1145/3506860.3506935

## Appendix 1

**Table 3.** The distribution and combination of sensors and modalities in the 36 reviewed MMCA studies

| Study | Sensors | | | | | | | | | | | | | Modalities | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | **Audio** | | **Visual** | | | | **Kinaesthetic** | **Tactile** | | **Physiological** | | |
| | Microphone | Camera | Kinect | Eye tracker | Manikin | EDA tracker | Heart rate tracker | Tang. Interface | Digital pen | Keyboard (software) | Breathing rate tracker | Accelerometer | GPS tracker | Prosodic | Semantic | Position | Gaze | Facial expression | Visible action | Spatial position | Logged action | Written output | Skin conductance | Heartrate | Respiration rate |
| Buckingham Shum et al., 2019 | 1 | 1 | | | 1 | 1 | | | | | | 1 | 1 | 1 | | | | | 1 | 1 | | | 1 | | |
| Chejara et al., 2020 | 1 | | | | | | | | | 1 | | | | 1 | | | | | | | | 1 | | | |
| Echeverria et al., 2019 | 1 | 1 | | | 1 | 1 | | | | | | 1 | 1 | 1 | | | | | 1 | 1 | | | 1 | | |
| Fernandez Nieto et al., 2022 | | | | | 1 | 1 | | | | | | | | | | | | | 1 | | 1 | | 1 | | |
| Fernandez Nieto et al., 2021 | | | | | 1 | 1 | | | | | | | | | | | | | 1 | | 1 | | 1 | | |
| Huang et al., 2019 | | | 1 | 1 | | 1 | | | | | | | | | | | 1 | | 1 | 1 | | | 1 | | |
| Kim et al., 2008 | 1 | | | | | | | | | | | 1 | 1 | 1 | | | | | | 1 | | | | | |
| Ma et al., 2022 | 1 | 1 | | | | | | | | | | | | 1 | 1 | 1 | 1 | 1 | | | | | | | |
| Madan et al., 2004 | 1 | | | | | 1 | | 1 | | | | 1 | | 1 | | | | | 1 | | 1 | | 1 | | |
| Malmberg et al., 2019 | | 1 | | | | 1 | | | | | | | | | | | | 1 | 1 | | | | 1 | | |
| Martinez-Maldonado et al., 2013 | 1 | | 1 | | | | | 1 | | | | | | 1 | | 1 | | | | | 1 | | | | |
| Martinez-Maldonado et al., 2020 | | | | | 1 | 1 | | | | | | 1 | 1 | | | | | | 1 | 1 | 1 | | 1 | | |
| Martinez-Maldonado et al., 2017 | 1 | | 1 | | 1 | | | 1 | | | | | | 1 | | 1 | | | | | 1 | | | | |
| Müller et al., 2018 | 1 | 1 | | | | | | | | | | | | 1 | | 1 | | 1 | | | | | | | |
| Nakano et al., 2015 | 1 | | | 1 | | | | | 1 | | | | | | 1 | | 1 | | | | | 1 | | | |
| Neubauer et al., 2016 | 1 | 1 | | | | | 1 | | | | | | | | 1 | | | 1 | | | | | | 1 | |
| Noroozi et al., 2019 | | 1 | | | | 1 | 1 | | | | | | | | | | | | 1 | | | | 1 | 1 | |
| Ochoa et al., 2018 | 1 | | 1 | | | | | 1 | | | | | | | 1 | | 1 | 1 | | | 1 | | | | |
| Olsen et al., 2020 | 1 | | | 1 | | | | | | 1 | | | | 1 | 1 | | 1 | | | | 1 | | | | |
| Ouhaichi et al., 2021 | 1 | 1 | | 1 | | 1 | 1 | | | | 1 | | | 1 | | | 1 | | | | | | 1 | 1 | 1 |
| Peng & Nagao, 2021 | 1 | 1 | | | | | 1 | | | | | | | 1 | 1 | 1 | | | | | | | | 1 | |
| Praharaj et al., 2018 | | | | | | | | | | | | | | 1 | | 1 | 1 | | | | | | | | |
| Reilly & Schneider, 2019 | 1 | | 1 | 1 | | 1 | | | | | | | | | 1 | 1 | 1 | | | | | | 1 | | |
| Scherer et al., 2012 | 1 | 1 | | | | | | | 1 | | | | | 1 | | | | | | | | 1 | | | |
| Schneider & Pea, 2015 | 1 | | | 1 | | | | | | | | | | | 1 | | 1 | | | | | | | | |
| Sinclair & Schneider, 2021 | 1 | | 1 | | | | | | | | | | | | 1 | 1 | | | | | | | | | |
| Spikol et al., 2017a | | 1 | | | | | | 1 | | | | | | | | 1 | 1 | | 1 | | 1 | | | | |
| Spikol et al., 2017b | 1 | 1 | | | | | | 1 | | | | | | 1 | | 1 | 1 | | | | 1 | | | | |
| Spikol et al., 2018 | 1 | 1 | | | | | | 1 | | | | | | 1 | | 1 | | | | | 1 | | | | |
| Stewart et al., 2021 | 1 | | | | | | | | | | | | | 1 | 1 | | | 1 | | | | | | | |
| Sturm et al., 2007 | 1 | 1 | | | | | | | | | | | | 1 | | 1 | | | | | | | | | |
| Viswanathan & VanLehn, 2018 | 1 | | | | | | | 1 | | | | | | 1 | | | | | | | 1 | | | | |
| Vrzakova et al., 2020 | 1 | | | | | | | | | | | | | | 1 | 1 | | | 1 | | | | | | |
| Worsley & Blikstein, 2014 | | 1 | 1 | | | 1 | | | | | | | | 1 | | | | | | 1 | | | 1 | | |
| Zhao et al., 2023 | 1 | | | | | | | | | | | | 1 | | 1 | | | | | 1 | | | | | |
| Zhao et al., 2022 | 1 | | | | | | | | | | | | 1 | 1 | | | | | | 1 | | | | | |
| **Total** | **27** | **17** | **8** | **6** | **6** | **12** | **4** | **8** | **2** | **2** | **1** | **5** | **6** | **21** | **12** | **13** | **11** | **6** | **11** | **8** | **12** | **3** | **12** | **4** | **1** |
| | Microphone | Camera | Kinect | Eye tracker | Manikin | EDA tracker | Heart rate tracker | Tang. Interface | Digital pen | Keyboard (software) | Breathing rate tracker | Accelerometer | GPS tracker | Prosodic | Semantic | Visible position | Gaze | Facial expression | Visible action | Spatial position | Logged action | Written output | Skin conductance | Heartrate | Respiration rate |

## Appendix 2

**Table 4.** Extraction table

| Extraction category | Description |
|---|---|
| Type of study | Type, e.g. Empirical research paper; Review paper; Conceptual paper; Methodological paper |
| Thematic focus and study rationale | Summary of the thematic focus and rationale for conducting MMCA study |
| Empirical setting | Country where study takes place; Educational level of participants; Discipline of participants |
| Conclusion/outcome | Brief summary of what the authors' main conclusion/outcome of the study is |
| Theory | Wider theory mentioned: e.g. cognitivist, socio-constructivist theory, management theory. |
| Relevant concepts or frameworks | Relevant concepts (and their definitions) or frameworks related to understand collaboration (e.g. collaborative problem solving (CPS), problem-based learning, team work) |
| Factors affecting collaboration/teamwork | What authors state as relevant factors affecting process and/or outcome of collaboration (e.g. group dynamics, linguistic coordination, coherence) |
| Sample | Sample from which MMCA data is collected |
| Sensors/devices | Sensors/devices used to capture the MMCA data |
| Mode/type of data used in MMCA | Mode/type of data that is collected |
| Variables/features inferred from data | Describe variable or features that authors infer from the data. They can be on the individual level (e.g. speaking time of individual participant) or on the group level (e.g. episodes of joint visual attention) |
| Aspects of collaboration indicated by variables/features | Describe aspects of collaboration that the authors claim to be indicated by the above variables/features separately or in combination. |