Chapter 6: Multimodal Learning Analytics - Rationale, Process, Examples, and Direction

Xavier Ochoa[1]

[1] Learning Analytics Research Network, New York University, New York, USA

DOI: 10.18608/hla22.006

ABSTRACT

This chapter is an introduction to the use of multiple modalities of learning trace data to better understand and feedback learning processes that occur both in digital and face-to-face contexts. First, it will explain the rationale behind the emergence of this type of study, followed by a brief explanation of what Multimodal Learning Analytics (MmLA) is based on current conceptual understandings and current state-of-the-art implementations. The majority of this chapter is dedicated to describing the general process of MmLA from the mapping of learning constructs to low-level multimodal learning traces to the reciprocal implementation of multimedia recording, multimodal feature extraction, analysis, and fusion to detect behavioral markers and estimate the studied constructs. This process is illustrated by the detailed dissection of a real-world example. This chapter concludes with a discussion of the current challenges facing the field and the directions in which the field is moving to address them.

Keywords: Multimodal, audio, video, data fusion, multisensor

The defining goal of Learning Analytics is the study of the low-level traces left by the learning process in order to better understand and estimate one or more learning constructs that are part of the process and, through carefully designed information tools, help the participants of that process to improve some desired aspects of it. The first works of Learning Analytics focused on the traces that were automatically generated when learners interacted with some type of digital learning tool. For example, Kizilcec, Piech, and Schneider [21] used the log of the actions performed by different groups of students in massive open online courses (MOOCs) to study course engagement, or Martin et al. [26] that use the low-level actions of students playing an educational video game study learning strategies. While these tools fulfill the goal of Learning Analytics, if we only focus on a single type of traces that are recorded in logs of digital tools, we risk oversimplifying the process of learning or even worse, misunderstanding the traces due to the lack of contextual information, two of the main critiques directed towards Learning Analytics from the educational research community [36].

The initial bias to base Learning Analytics works solely on the data of interactions of students with digital learning tools can be explained by the relative abundance of this type of data. Digital tools, even if not initially designed with analytics in mind, tend to automatically record, in fine-grained detail, the interactions with their users. The data describing these interactions is stored in many forms; for example, log-files or word-processor documents can be later mined to extract the traces to be analyzed. Also, the low technical barriers to process this type of data make digital the ideal place to start Learning Analytics research. On the other hand, in learning processes that occur without the intervention of digital tools, for example, face-to-face blackboard-based collaborative problem solving, the actions of learners are not automatically recorded. Even if some learning artifacts exist, such as student-produced physical documents or photographs, they need to be converted before they can be processed. Without traces to analyze, computational models and tools used traditionally in Learning Analytics are not applicable.

The existence of this bias towards learning contexts where digital tools are the main form of interaction could produce a streetlight effect [17] in Learning Analytics. The streetlight effect consists of looking for solutions where it is easy to search, not where the real solutions are most probable to be found. Translating this effect to Learning Analytics, it to use a given learning trace, for example, access to materials on the LMS, to estimate a learning construct, for example, engagement, just because we only have access to that data, not because we have a theoretically or empirically strong indication that level of access is a robust predictor of engagement. A more holistic analysis of even the simplest learning construct requires the examination of different sources of evidence at different levels of complexity. For example, a human instructor trying to assess the level of engagement of students could review not only their online actions but their participation in face-to-face activities, their academic and social interactions with others, the quality of their work, and even their body language during lectures. Even if no single dimension

independently is a very robust indicator of the desired construct, the triangulation between different but related and complementary sources of information is bound to provide stronger evidence upon which an intervention decision could be taken with confidence [30].

Addressing the streetlight effect in Learning Analytics requires that, instead of being guided by the data that is available, the study start with theory- or experience-based analysis of how the desired learning construct manifest itself through behavioral markers in different contexts and identifying what low-level traces can be used as evidence of those behaviors. Then, technological solutions need to be found to record the learning process in the context where it occurs and extract the identified traces. Finally, these traces need to be analyzed and fused to detect the behavioral markers and finally to robustly estimate the learning construct of interest and feedback the information to the participants of the learning process in an understandable and actionable way. The nascent subfield of Multimodal Learning Analytics (MmLA) strives to fulfill this tall request. This chapter is an initial guide for researchers and practitioners who want to explore this sub-field. It will discuss in detail the MmLA focus of study, its processes, and current examples of how it instantiates in real-world scenarios.

1 WHAT IS MULTIMODAL LEARNING ANALYTICS

In its communication theory definition, multimodality refers to the use of diverse modes of communication (textual, aural, linguistic, spatial, visual, et cetera) to interchange information and meaning between individuals [23]. It is different from the concept of multimedia, using diverse media to communicate information. The media — movies, books, web pages, or even air — are the physical or digital substrate where a communication mode can be encoded. Each mode can be expressed through one or several media. For example, speech can be encoded as variations of pressure in the air (in a face-to-face dialog), as variations of magnetic orientation on a tape (in a cassette recording), or as variations of digital numbers (in an MP3 file). As well, the same medium can be used to transmit several modes. For example, a video recording can contain information about body language (posture), emotions (face expression), and tools used (actions).

Multimodal Learning Analytics is rooted in the Multimodal Interaction Analysis framework (Norris, 2020) that exhort the integration of multimodal information (human verbal and non-verbal forms of communication together with information about the objects used as part or medium of the communication and the contexts in which this communication occurs) to better study and understand how humans act and interact with others, with technology, and with the environment. Translating this framework to educational settings, Paulo Blikstein first formally introduced the concept of Multimodal Learning Analytics at the 3rd Learning Analytics and Knowledge Conference (LAK) 2013 in a homonymous paper [5]. In this paper, MmLA is defined as "a set of techniques that can be used to collect multiple sources of data in high frequency (video, logs, audio, gestures, biosensors), synchronize and code the data, and examine learning in realistic, ecologically valid, social, mixed-media learning environments." Unpacking this definition, we can observe the three main operative processes of MmLA, already hinted in the introduction of this chapter: use of diverse sources of learning traces (multimodal data), processing and integration of these traces (multimodal analysis and fusion), and the study of human behavior in real learning environments (learning behavior detection and learning construct estimation).

While the term Multimodal Learning Analytics was formally coined in 2013, the application of the Multimodal Interaction Analysis framework to educational context has always been part of the Learning Analytics agenda. Already in the first LAK conference, [6] proposed its use in the then-nascent field. Before LAK, what can now be considered bonafide MmLA works were published at the International Conference for Multimodal Interaction (ICMI), which hosted the 1st Multimodal Learning Analytics workshop already in 2012 [34]. However, the idea of using different communication modalities to study learning predates even the terms Multimodal Interaction and Learning Analytics and it is common in traditional experimental educational research. In this research tradition, a human observer, which by nature is a multimodal sensor, is tasked with noting and annotating relevant interactions that occur in real-world in-the-wild learning contexts for further qualitative analysis [18]. Technologies such as video and audio recording and coding and tagging tools have made this observation less intrusive and more quantifiable [9, 25]. MmLA, however, presents several important differences with traditional educational research practices: 1) In MmLA, the collection of the data is performed by low-cost high-definition sensors that enable the capture of the traces with a level of detail that was not feasible before, 2) in MmLA, early coding happens automatically through the use of machine learning and artificial intelligence algorithms, eliminating the limits in both the number of codes and the time length that is imposed by the manual nature of human coding. 3) In MmLA, the analysis and fusion of the data can be (semi-) automated providing systems that could be used in real- or near real-time and, 4) in MmLA, the result of the analysis is not only used to expand our understanding of the learning process being observed but could also be used to create an analytic tool to provide information back to students and/or instructors to generate a feedback loop to improve learning as it is happening. While both traditional multimodal educational research and MmLA share a common interest in the different ways in which humans interact during learning activities, the affordances provided by the speed and scale of MmLA open a different set of opportunities to understand and improve learning processes.

A good way to understand the kind of opportunities that MmLA affordances provide is to review some of the most notable examples of this sub-field available in the literature. Table 1 presents a non-exhaustive list of exam-

ples of successful applications of MmLA techniques in diverse learning settings. The list mentions the different modalities used in the work and the learning construct being studied or estimated. As it can be seen in the table, MmLA has been used in contexts as dissimilar as traditional classrooms to medical simulations and educational games. While a great variety of modes are explored video- and audio-based modes such as gaze, movement, gestures, and speech are the most common, followed by bio-signals (mental activity and electrodermal activity). However, depending on the circumstances specialized modes are used (pen strokes for calligraphy and manikin interactions in medical simulations). The variety of learning constructs being studied is even more diverse than the learning settings, exemplifying the great flexibility of MmLA as a research and practice tool. Di Mitri, Schneider, Specht, and Drachsler [13] can provide the reader with a wider and deeper review of existing MmLA systems together with their modalities and investigated constructs. While all the systems in Table 1 and the ones mentioned in Di Mitri et al. have different objectives and implementations, they all follow a similar process. This high-level MmLA process will be explained in the next section.

2 THE PROCESS OF MMLA: FROM CONSTRUCT TO TRACES AND BACK AGAIN

Due to its nature, most of MmLA studies and tools, even if it is not explicit in their published description, follow a common process. This process can be roughly divided into two reciprocal phases: Mapping and Execution. During the mapping phase, a logical path is found between theoretical learning constructs of interest and multimodal data traces that can be observed during the learning process. During the execution phase, that path is reversed and extracted multimodal data traces are used to estimate the desired learning constructs. While the second phase, execution, receives a great deal of attention due to its technical complexity, it is the first phase, mapping, where MmLA directly tackles the streetlight effect problem in Learning Analytics. The following subsections will explain the different steps inside these two phases together with the main concerns that emerge with the use of multimodal data.

2.1 Mapping Phase: From Learning Constructs to Multimodal Data Traces

Thanks to some of its roots in Experimental Psychology and Educational Research, Learning Analytics have adopted the idea of a construct, most commonly referred to as a learning construct, to organize and explain the reason behind the measurements, analysis and interventions conducted [11]. A learning construct can be defined as a concept or idea related to students' behaviors, attitudes, learning processes and experiences. By definition, a construct is not directly observable or measurable but manifests itself through behaviors that occur when the learner interacts with the learning environment. Those behaviors can then be used to estimate the value, graduation, or intensity of the construct. For example, intelligence is a common construct used in education. To be able to estimate the intelligence of individuals, we expose them to situations where their need to use their complex cognitive abilities, for example exposing them to a set of complex problems, puzzles, or an IQ test and using the time and number of correct answers to estimate how intelligent they are. The mapping phase has four steps and results in a tree-like map that links the learning construct of interest with the observable data traces. Figure 1 presents a detailed view of this tree, while Figure 2 shows this phase as a part of the MmLA process. This mapping process is not unique to MmLA and has been proposed initially by Worsley et al. [41] and refined by Echeverria [14]. However, this model is especially well suited for studies that involve multimodal data.

The first step in the mapping phase is the definition of the learning construct of interest. This selection is ideally guided by the needs of the learning process stakeholders as discovered by the researcher but sometimes is determined by the interest or curiosity of the researcher. The initially selected construct could encompass a large set of diverse behaviors, for example, "collaboration skills". In this case, we could divide the learning construct into sub-constructs. We can divide the "collaboration skills" construct into "participation" and "active listening" sub-constructs each one capturing a different subset of the behaviors connected to collaboration skills.

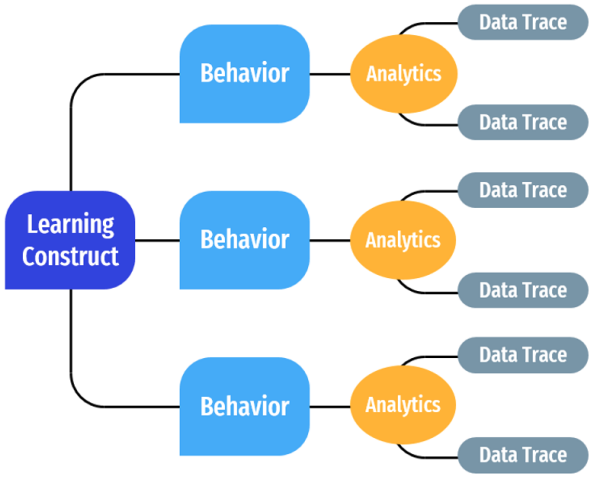

Figure 1: Construct Mapping detail tree-structure, adapted from [14].

2.2 Execution Phase: From Multimodal Data Traces to Learning Constructs

Once the mapping between Learning Constructs and low-level multimodal data traces is complete (at least as a first draft in the mind of the researcher or practitioner), a Multimodal Learning Analytics System can be built. In general, this system could have two different goals. The first one is research-oriented and starving to generate new gen-

Table 1: Non-exhaustive list of examples of the application of MmLA system in different learning settings.

Learning Setting	Reference	Main Multimedia Data	Main Learning Construct
Calligraphy Learning	[24]	Gaze location on screen (eye-tracking), pen strokes, movement	Mental effort
Classrooms	[32]	Gaze direction (eye-tracking), mental activity (EEG), movement, subjective view (video), subjective hearing (audio)	Classroom orchestration
Collaborative Problem Solving	[15]	Touch coordinates, speaking time, participant hand position	Contribution to solving the problem
Dance	[33]	Facial expression, gaze, posture, movement	Dance skills
Educational Games	[19]	Keystrokes, mental activity (EEG), Gaze location on screen (eye-tracking), facial expression (video), electrodermal activity (EDA)	Learning gains
Embodied Cognition	[2]	Gaze, gestures, movement	Concept understanding
Intelligent Tutoring Systems	[20]	Scores, time on task, number of tasks, speech pauses and length	Affect
Making	[40]	Human video coding, skeletal tracking	Efficacy of learning practices
Medical Simulation	[27]	Interactions with a patient manikin, use of digital checklist, location, speech	Team collaboration
Oral Communication	[35]	Posture, gestures, speech volume and cadence	Oral presentation skill
Programming	[10]	Usage of digital system, speech	Collaboration and communication

eralizable knowledge about the learning construct. For example, what are the main differences between the engineering building processes of novices and experts [42]. The second could be practice-oriented, striving to provide an analytic tool to improve the learning process for the participants. For example, an automated feedback system to improve oral presentation skills [29]. While these two objectives are not necessarily mutually exclusive, MmLA works tend to align with one or the other due to implementation requirements that will become apparent when this phase is discussed in detail.

The execution phase can be seen in the second lower part of Figure 2. It runs in reverse order compared to the mapping stage and consist usually of four steps. First, multimedia signals are recorded from the relevant participants in the learning activity. Then, these recordings are automatically processed to extract low-level multimodal data traces. These low-level traces are then (semi-) automatically analyzed and fused to produce high-level traces. These high-level traces are used to detect the occurrence of desired behaviors and to estimate the studied learning (sub-)constructs. Finally, if the final goal of the system is to build an analytic tool, the obtained estimations are used to feed the tool providing the information back to the learning process participants. The following subsections will present the requirements and operation of these steps in detail.

2.3 Multimedia Recording

The first step in the execution phase is to be able to register or record all the relevant signals that contain the data traces identified in the mapping phase. In the case of the interactions of digital tools, this capture could be as simple as adding a logging statement in relevant parts of the tool's code. On the other hand, in situations that require the capture of non-computer-mediated actions, such as a face-to-face conversation between two individuals, the use of different types of sensors is needed. These sensors could be as simple as a webcam or as sophisticated as a magnetic resonance imaging (MRI) machine. Moreover, the multimodal aspect of MmLA systems usually requires the use of several sensors, each one specialized in a different type of media. For example, a webcam for video, a microphone for audio, a digital pen for the learner's notes. There is a large range of sensors and modalities that have been used in MmLA systems [13]. While the selection of the right type of sensors and the design and setup of the recording apparatus is an engineering problem, researchers and practitioners alike should be aware of the affordances, limitations, and scalability of these components to create effective MmLA systems.

2.4 Multimodal Feature Extraction

Once the raw multimedia data is captured, the next step is to extract the identified multimodal data traces embedded in those recordings. This extraction, in general, requires a computer algorithm that can process the raw recording or data file and isolate or generate the trace for the required modality. For example, if we require the body posture of the participants and we have a video recording, we can use computer vision algorithms, more specifically Convolutional Pose Machine [39], for example, that implemented in OpenPose [7], to obtain the position of the skeletal joints and pose of all the individuals present in each video frame. In another example, speech to text algorithms, for example, the one provided as a service by Google Speech, can be used to extract the verbal content of the audio signal recorded by a microphone. Similar to the recording step, while it is not necessary to possess full knowledge of how each extraction algorithm operates, it is highly recommended that researchers or practitioners understand the affordances and limitations of those algorithms.

2.5 Multimodal Analysis and Fusion

The traces extracted from raw data are defined for a single modality. For example, feature extraction might compute student eye gaze direction or voice pitch. While there are some cases in which low-level unimodal traces are enough to estimate the desired behavior, most commonly these traces need to be processed and fused together to create higher-level traces that are more accurate and robust predictors. For example, if the behavior of joint visual attention in a collaborative activity around a table is of interest, the estimated individual gaze direction from each participant has to be fused together with the direction of the other participant's gaze to detect if two or more of them intersect inside a given region in the table. In another example, turn-taking information can be extracted from the change in the current speaker trace. In a more complex example, turn-taking information, paired with idea identification information obtained from speech, could be used to identify idea uptake traces. The development of these fusion algorithms is still an open challenge in MmLA and very much guided by the analytic description during the mapping phase. The recommended approach to tackle the construction of these algorithms is to develop a human rubric to measure as objectively and reproducible as possible the observation of the high-level traces, then using a mixture of theoretical knowledge and Principal Factor Analysis to select promising low-level traces to model the desired high-level one. This technique is explored in Chen, Leong, Feng, and Lee [8].

2.6 Behavior Detection and Construct Estimation

This step in the execution phase is not particularly different for MmLA when compared with more traditional works in Learning Analytics and Educational Research. Once the results of the analysis and fusion phase provide information about the occurrence of the identified behaviors, computational or statistical analysis (or qualitative analysis in the case of research-oriented MmLA systems) can be used to estimate the level, grade, or intensity of the studied learning (sub-)construct(s). The only main consideration for MmLA systems is the increased level of uncertainty in the detection of behavioral markers. In a similar way in which the estimation of inter-rater coefficients is used to assess the reliability of the coding of the

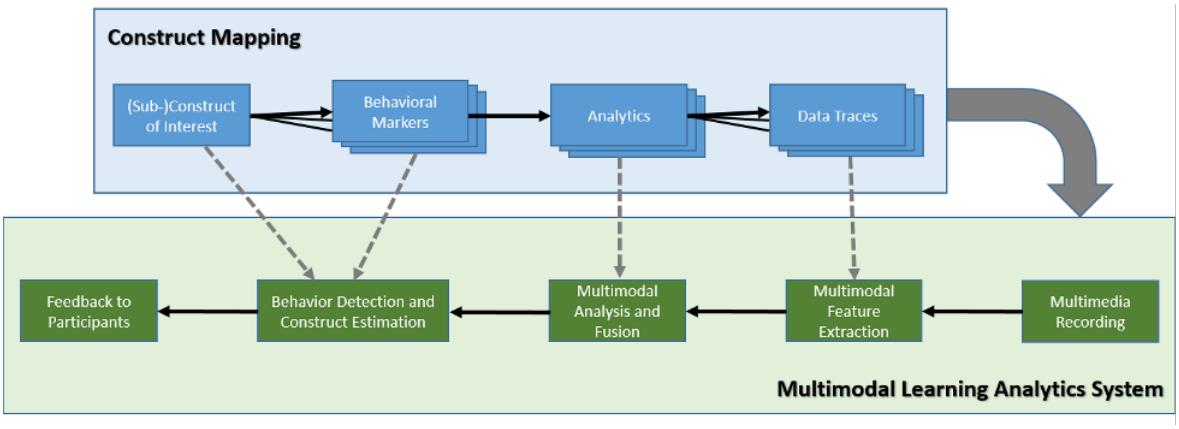

Figure 2: Diagram of the MmLA Process.

ground truth, the measured accuracy of the automated detection should be calculated against one or more human coders. If this a research-oriented MmLA system, this is the final step in the process. The estimation of the construct(s) can be used to draw generalizable conclusions about the nature, workings, or efficiency of the learning process, and through the publications of these results, improve the general knowledge about how humans learn and maybe improve new designs of the studied or similar learning process.

2.7 Feedback to Participants

If the goal of the system is to provide reflection opportunities and actionable feedback to the participants of the learning an analytic tool has to be built and fed with the data generated during previous steps. For this kind of tool to be effective, it has to consider what information to present, when to present it and how to present it [22]. For instance, letting a teacher know that a group was struggling after the activity has been completed is less effective than letting them know during the activity when there is the possibility to intervene. Notwithstanding, there may be instances where it is best not to intervene, as well as situations where instructors wish to reflect on how their prompts impacted student-student collaboration. Switching to the student perspective, it might be the case that providing each student with a dashboard presenting several collaboration-related measurements in their smartphones during the activity could distract them from the activity itself. The information provided by MmLA systems enables the exploration of new and innovative ways to close the loop of Learning Analytics.

Multimodality embedded in the system can be used to create more natural ways to provide the right information, in the right moment and in the right modality. These multimodal interfaces predate MmLA but have been described in other research communities. As an example, Alavi and Dillenbourg [1] successfully tested ambient signaling lights to support teachers to easily identify struggling groups during supervised collaborative problem-solving. Bachour, Kaplan, and Dillenbourg [4] experimented with the use of an illuminated interactive tabletop to provide real-time feedback to students about their participation in the conversation.

3 MMLA PROCESS IN ACTION

To demonstrate how the diverse steps of the MmLA process are implemented, a real MmLA study will be dissected and analyzed. This study is a representative of one of the oldest and widest applications of MmLA, providing feedback for oral presentations [29, 29].

3.1 Oral Presentation Feedback System

This example describes a multimodal system for automated feedback for oral presentation skills [28, 29]. This system was designed and implemented in a mid-size polytechnic higher education institution on the coast of Ecuador. In a nutshell, this system allows students to practice oral presentations in front of a recorded audience and to receive a report that indicated if they made common presentations errors such as looking at the slides for long periods or speaking too softly. Figure 3 present the physical layout of the system. The following subsections will describe the MmLA process followed in the implementation of this tool.

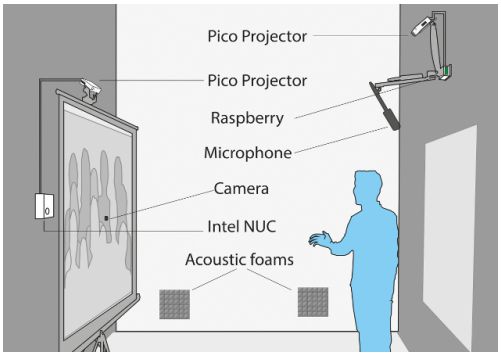

Figure 3: Physical layout of the multimodal system for oral presentation feedback, taken from [28].

3.2 Construct Mapping Phase

Figure 3 presents the construct mapping for this first example. The main objective of this tool was to help learners to develop basic oral presentation skills. By consultation with communication professionals, the "Basic Oral Presentation Skill" construct was connected with four observable behaviors: 1) Looking at the audience; 2) Maintaining an open posture; 3) Speaking loudly; and 4) Avoiding filled-pauses. The next step in the mapping was to identify the analytics to detect the behaviors. For example, looking at the audience can be detected when the gaze of the presenter was directed towards the camera (that was embedded in the middle of the recorded audience projection. In another example, the presence of filled pauses ("ahh", "umm", among others) was detected by an analysis of variance of speech formants. Finally, the multimodal data traces needed for each analytic was extracted. In this case, each analytic is connected to just one trace. In total, four traces need to be extracted: gaze, posture, speech volume, and speech formants. This mapping is very simple, there no triangulation for behavioral detection, there is no multimodal fusion strategies. A consequence of this design is that the accuracy of the feature extraction needs to be high in order to avoid behavior misidentification.

3.3 Execution Phase

The first step of the execution phase was to determine the sensors needed for the multimedia recording. It was determined that gaze and posture could be extracted from a video feed of the presenter recorded by a webcam embedded in the middle of the screen where the recorded audience was projected. Alternatively, a hardware depth sensor, such as Microsoft Kinect could have been used to extract these to modalities, but a camera was preferred due to implementation cost, leaving the heavy processing for a centralized software implementation. The speech volume and speech formats were capture in the audio signal recorded by a mono-channel microphone located above the presenter.

For the multimodal feature extraction step, diverse software libraries were used. For the posture, OpenPose, a convolutional pose machine, was used to obtain the 2D position of the skeletal joints. Using part of the skeletal joints the head posture (relative position of ears, nose, and neck) was calculated as a proxy of gaze, given that the video quality was not enough to perform a landmark analysis of the face. Given that only a coarse gaze direction is needed (looking at the audience, looking away from the audience) was needed, this setup was determined be a good compromise between precision and implementation cost. For the speech features (volume and formants), a commonly used software package for analysis of speech characteristics (PRAAT) was employed. The accuracy of the extraction of these characteristics was performed [29] and was determined to be sufficient for the application at hand. The multimodal analytics and fusion step was straightforward given the lack of any fusion between features. For the detection of an open posture, the random forest model was trained with human coded images of open and close postures, mostly related to the position of the arms with respect to the body, especially the hips. This model was then used to classify the postures as open or closed. In the case of volume, a simple threshold detector was used to differentiate between loud and soft speech.

The detection of the behaviors was also straight-forward. An error rate approach was used to provide a value to how much a given behavior was observed. For example, the percentage of time that presenter keeps their gaze looking towards the projected audience versus away from it. These percentages were used then to calculate a score (based on recommendations by the original communication professionals). These scores were linearly added to estimate the level of oral presentation skills in the participant.

Finally, the calculated scores, together with the information generated through the whole execution phase was used to create a multimedia feedback report (Figure 5). This report presented the final score together with the scores for each one of the behaviors. The presenter was also able to see or hear recordings of good and bad examples of each of the scored behaviors.

4 CHALLENGES & DIRECTIONS

It is the intention of this chapter to introduce the sub-field of MmLA, its process, its potentialities, and to provide examples of its state-of-the-art. However, no discussion about MmLA is complete without addressing the multiple methodological, technical, practical, and ethical challenges that it currently confronts and how the MmLA community is trying to address them moving forward.

4.1 Methodological Challenges

One of the most pressing issues that MmLA, as a field, faces is the lack of homogeneous methodological approaches and a compendium of best practices. Due to the novelty of the field, which is the intersection point of several research traditions (multimodal interaction, educational research, artificial intelligence, among others), each study uses different approaches for the validation of its measurements, fusion of multimodal information, and even the definition of constructs, behavioral markers, analytics, and modalities. This complete diversity, while initially beneficial as a way to explore the affordances and limitations of the field, it is now generating problems in the generalization, reproducibility, and sharing of results. It also limits the capacity of MmLA to contribute to a common theoretical body-of-knowledge as each study is a one-off enterprise.

The need to share definitions, methods, and best practices was early identified by the community. The first MmLA workshop was already organized in 2012 [34] and has been repeated yearly since. The MmLA community has also formally created a Special Interest Group (SIG) inside the Society for Learning Analytics Research (SoLAR). All these efforts have started to bear fruit in recent years as several publications have started to catalog and sys-

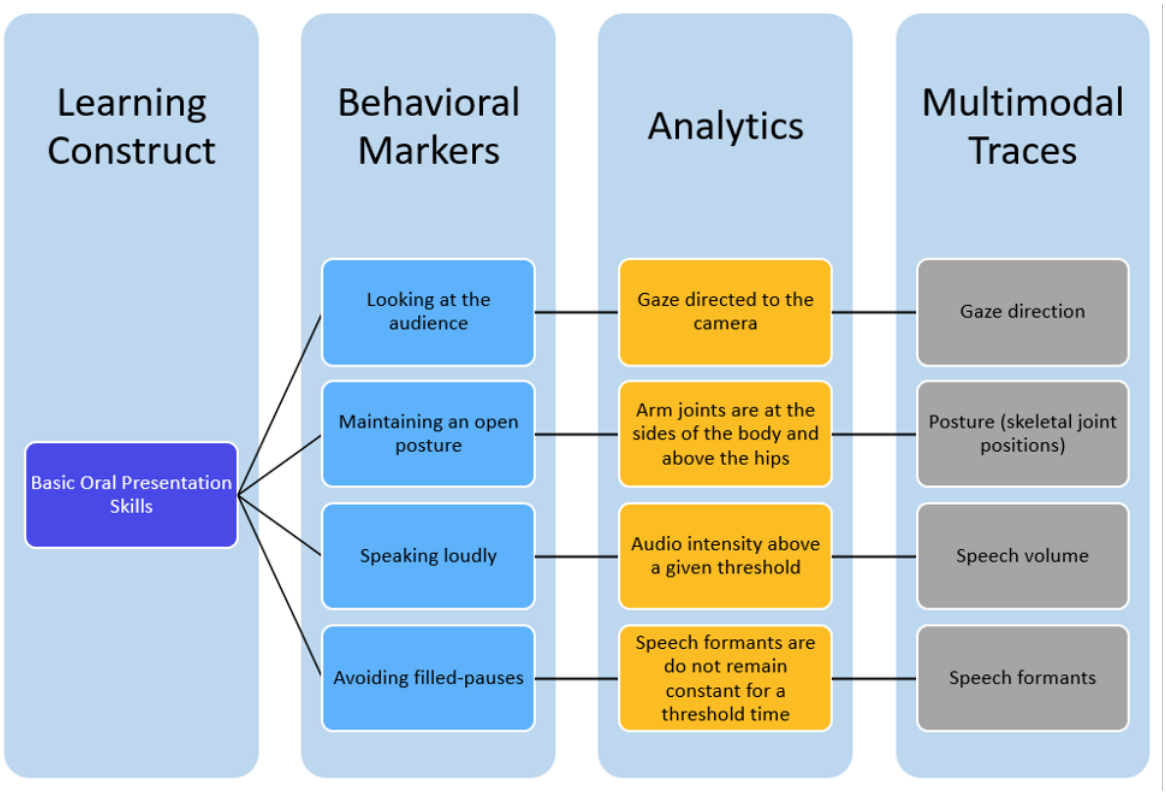

Figure 4: Construct Mapping for the Oral Presentation Feedback Tool.

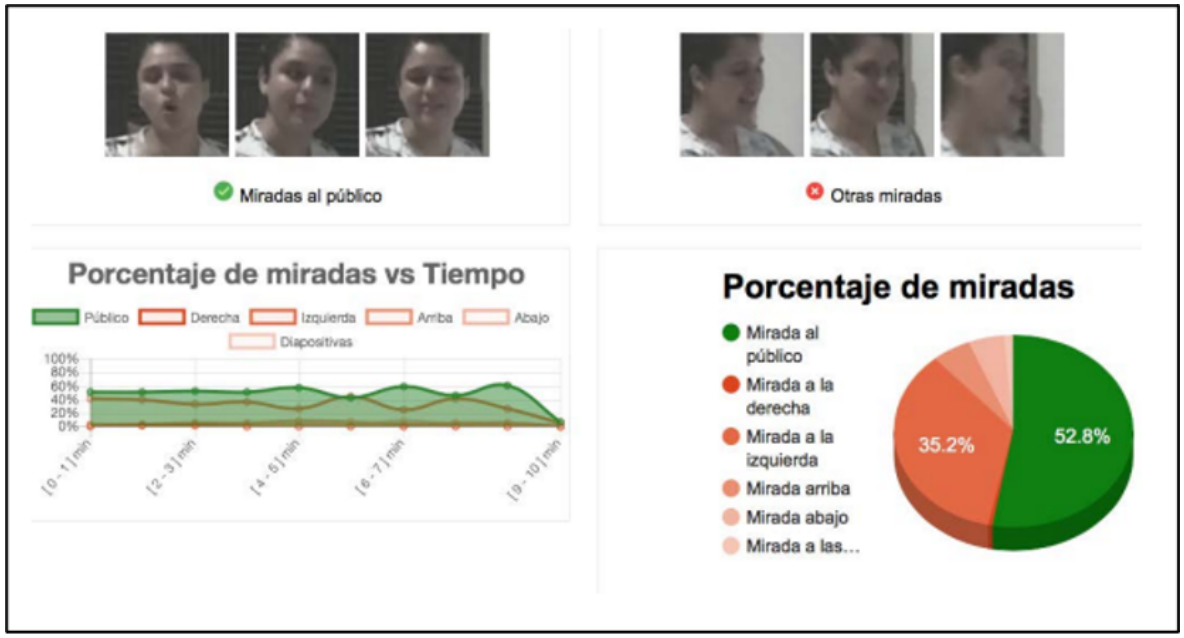

Figure 5: Example the multimedia report from the oral presentation feedback tool, taken from [29].

tematize the different approaches used by MmLA work and proposing common conceptual and methodological frameworks to better align the different research traditions inside MmLA. Examples of this new wave of integrative research are the frameworks proposed by Worsley et al. [41], Eradze, Rodríguez-Triana, and Laanpere [16], Di Mitri et al. [13], Sharma, Papamitsiou, and Giannakos [37], and Echeverria [14]. This last example has been used as a base for the MmLA construct mapping process presented in this chapter. It is expected that in the following years, these frameworks will provide a common ground for MmLA works to be more comparable, generalizable, and incrementally improved by others outside their original creator team.

4.2 Technical Challenges

Another aspect that hinders a more accelerated progression of MmLA is the technical difficulty that implementing multimodal analytic systems entails. While MmLA benefits from state-of-the-art developments in sensor technologies, digital signal processing, machine learning, and artificial intelligence in general, it also requires technical experts in these areas to be involved in the design and implementation of MmLA systems. Technical issues raised by the distributed operation of the sensors, synchronization of the signals, advanced feature extraction, and multimodal fusion strategies keep most educational-focused teams, without access to those experts, away from exploring MmLA solutions to study real-world learning processes. This is a problem shared by the Multimodal Interaction community in general. Tentative technical solutions have started to emerge in germane fields. For example, Social Signal Interpretation (SSI) framework [38] provide a software framework that offers connection with a wide variety of sensor, warrantied synchronization even with sensors distributed across a network, machine learning model training and use, multimodal fusion and behavior detection. While not easy-to-use by any metric, this is a step in the direction of simplifying the design and implementation of MmLA systems. Another emerging, but not currently widely tested, software framework available is the Microsoft Platform for Situated Intelligence [3] that promises a more robust set of development and visualization tools. It is expected that in the immediate future the construction of MmLA systems to be greatly facilitated by this kind of software solutions that remove the need to pay close attention to the technical details and facilitate the researchers to concentrate on the study of the learning process.

4.3 Practical Challenges

Most of the current MmLA tools only reach the prototype stage [12]. While useful for research on MmLA and its potential, these tools have almost no impact on real-world learning processes. To bridge the gap between an interesting technical prototype and a pedagogically-integrated solution, MmLA, as a field, need to pay more attention to practical issues that affect the attractiveness of the MmLA systems for educators and educational institutions. The most important of these issues are cost (initial cost and maintenance cost), easy-of-use (no technician should be required for day-to-day use), robustness (the system should graciously manage hardware, network, or software problems), and scalability (it should be feasible to deploy the system institution-wide). These are common requirements for any learning technology, including any Learning Analytics tool. However, solving these practical problems is beyond the interest and knowledge of most researchers, requiring stronger participation of learning technology practitioners that seeing the potential of MmLA translate the prototypes into solutions that can be easily deployed in-the-wild. Ochoa and Dominguez [28] offers an example of a MmLA tool that was successfully implemented in-the-wild.

Ethical challenges are the “elephant-in-the-room” for MmLA. Not so much because it is not spoken about (they are a constant theme of debate among MmLA researchers and practitioners) but because they generate issues that can overweight any methodological, technical, or practical consideration. Capturing interaction information with digital tools already raises privacy concerns among students and instructors [31]. The installation and use of recording systems that technically mimic (and sometimes exceed) "1984" levels of surveillance are bound to meet understandable strong resistance from the learning process stakeholders, especially those under observation. While these issues are less problematic for research-oriented MmLA systems used in laboratory settings, they can completely block even the idea of using them in real learning environments.

The main way in which the MmLA community is trying to address these challenges is by clearly separating research from practice. The data captured in research-oriented MmLA systems in-the-lab, after the required consent forms are signed, could be used to advance the state of the knowledge in the field with just the minimum required safeguards for the privacy of the participants and their immediate benefit. The data produced in these settings usually belongs and is controlled by the research team that built the tool. On the other hand, data produced by a practice-oriented MmLA system in-the-wild can only be used for the immediate benefit of the observed participant. Also, the data belongs and its use and storage should be controlled by the participant. Strong safeguards should be in place to deter the use of this data for something different than its original purpose to feedback the learning process participants. Only with these safeguards, practitioners should be able to address natural negative perceptions of technology that could be misused for unduly monitoring and surveillance.

5 CONCLUSION

Learning Analytics has revolutionized the way in which we study and try to improve learning processes. However, its initial bias towards studies and tools involving only computer-based learning contexts jeopardizes its applicability and conclusions for learning in general. The MmLA

strives to widen the horizons of Learning Analytics, including richer and possible more relevant sources of data and including also learning context to which traditional Learning Analtyics could not be applied due to the lack of pre-existing data. As it can be inferred from the discussions in this chapter, especially for the current challenges and directions, MmLA is still young with many issues to be addressed. However, it is also a fast-growing and connected community of researchers and practitioners in constant search of innovative solutions to those issues. This community is also showing strong signs of maturing, such as the recent proposal of methodological frameworks integrating learning theories and multimodal interaction analysis and lowering the technological barriers of entry. This chapter, apart from being an introduction to MmLA, is an invitation for existing Learning Analytics researchers and practitioners to explore the use of multiple modalities in their own studies and tools. The MmLA community will openly share its knowledge, methodologies, code, successes, and failures. While current MmLA is considered a sub-field of Learning Analytics, is the belief of the author that in the future most of Learning Analytics studies will be multimodal in nature as learning itself is.

REFERENCES

[1] Hamed S. Alavi and Pierre Dillenbourg. "An ambient awareness tool for supporting supervised collaborative problem solving". In: *IEEE Transactions on Learning Technologies* 5.3 (2012), pp. 264–274. DOI: 10.1109/TLT.2012.7.

[2] Alejandro Andrade. "Understanding student learning trajectories using multimodal learning analytics within an embodied-interaction learning environment". In: *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM, 2017, pp. 70–79. ISBN: 1-4503-4870-X.

[3] Sean Andrist, Dan Bohus, and Ashley Feniello. "Demonstrating a framework for rapid development of physically situated interactive systems". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Mar. 2019, pp. 668–668. ISBN: 2167-2148. DOI: 10.1109/HRI.2019.8673067.

[4] Khaled Bachour, Frederic Kaplan, and Pierre Dillenbourg. "An interactive table for supporting participation balance in face-to-face collaborative learning". In: *IEEE Transactions in Learning Technology* 3.3 (2010), pp. 203–213. ISSN: 1939-1382. DOI: 10.1109/tlt.2010.18.

[5] Paulo Blikstein. "Multimodal learning analytics". In: *Proceedings of the Third International Conference on Learning Analytics and Knowledge*. Leuven, Belgium: Association for Computing Machinery, 2013, pp. 102–106. DOI: 10.1145/2460296.2460316. URL: https://doi.org/10.1145/2460296.2460316.

[6] Paulo Blikstein. "Using learning analytics to assess students' behavior in open-ended programming tasks". In: *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. Banff, Alberta, Canada: Association for Computing Machinery, 2011, pp. 110–116. DOI: 10.1145/2090116.2090132. URL: https://doi.org/10.1145/2090116.2090132.

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields". In: *IEEE transactions on pattern analysis and machine intelligence* 43.1 (2019), pp. 172–186. ISSN: 0162-8828.

[8] Lei Chen, Chee Wee Leong, Gary Feng, and Chong Min Lee. "Using multimodal cues to analyze MLA'14 oral presentation quality corpus: Presentation delivery and slides quality". In: *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*. 2666640: ACM, 2014, pp. 45–52. DOI: 10.1145/2666633.2666640.

[9] Paul Cobb, Jere Confrey, Andrea diSessa, Richard Lehrer, and Leona Schauble. "Design experiments in educational research". In: *Educational Researcher* 32.1 (Jan. 2003), pp. 9–13. ISSN: 0013-189X. DOI: 10.3102/0013189X032001009. URL: https://doi.org/10.3102/0013189X032001009 (visited on 12/23/2020).

[10] Hector Cornide-Reyes, René Noël, Fabián Riquelme, Matías Gajardo, Cristian Cechinel, Roberto Mac Lean, Carlos Becerra, Rodolfo Villarroel, and Roberto Munoz. "Introducing low-cost sensors into the classroom settings: Improving the assessment in agile practices with multimodal learning analytics". In: *Sensors* 19.15 (2019), p. 3291.

[11] Lee J. Cronbach and Paul E. Meehl. "Construct validity in psychological tests". In: *Psychological Bulletin* 52.4 (1955), pp. 281–302. ISSN: 1939-1455(Electronic),0033-2909(Print). DOI: 10.1037/h0040957.

[12] Mutlu Cukurova, Michail Giannakos, and Roberto Martinez-Maldonado. "The promise and challenges of multimodal learning analytics". In: *British Journal of Educational Technology* 51.5 (2020), pp. 1441–1449. ISSN: 0007-1013. DOI: 10.1111/bjet.13015. URL: https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13015.

[13] Daniele Di Mitri, Jan Schneider, Marcus Specht, and Hendrik Drachsler. "From signals to knowledge: A conceptual model for multimodal learning analytics". In: *Journal of Computer Assisted Learning* 34.4 (Aug. 2018), pp. 338–349. ISSN: 0266-4909. DOI: 10.1111/jcal.12288. URL: https://doi.org/10.1111/jcal.12288 (visited on 12/23/2020).

[14] Vanessa Echeverria. "Designing Feedback for Collocated Teams using Multimodal Learning Analytics". PhD thesis. University of Technology Sydney, 2020.

[15] Vanessa Echeverria, Roberto Martinez-Maldonado, Katherine Chiluiza, and Simon Buckingham Shum. "DBCollab: Automated feedback for face-to-face group database design". In: *Proceedings of the 25th International Conference on Computers in Education, ICCE 2017-Main Conference Proceedings*. 2017. ISBN: 986-94012-6-0.

[16] Maka Eradze, Maria Jesus Rodriguez Triana, and Mart Laanpere. "How to aggregate lesson observation data into learning analytics datasets?" In: *6th Multimodal Learning Analytics (MMLA) Workshop and the 2nd Cross-LAK Workshop co-located with 7th International Learning Analytics and Knowledge Conference (LAK 2017)*. 2017.

[17] David H. Freedman. "Why scientific studies are so often wrong: The streetlight effect". In: *Discover Magazine* 26 (2010).

[18] Meredith Damien Gall, Walter R. Borg, and Joyce P. Gall. *Educational research: An introduction, 6th ed*. Educational research: An introduction, 6th ed. White Plains, NY, England: Longman Publishing, 1996. ISBN: 0-8013-0980-8 (Hardcover).

[19] Michail N. Giannakos, Kshitij Sharma, Ilias O. Pappas, Vassilis Kostakos, and Eduardo Velloso. "Multimodal data as a means to understand the learning experience". In: *International Journal of Information Management* 48 (2019), pp. 108–119. ISSN: 0268-4012.

[20] Ruth Janning, Carlotta Schatten, and Lars Schmidt-Thieme. "Multimodal Affect recognition for adaptive intelligent tutoring systems". In: *International Journal of Information Management*. 2014.

[21] René F. Kizilcec, Chris Piech, and Emily Schneider. "Deconstructing disengagement: analyzing learner sub-populations in massive open online courses". In: *Educational Data Mining Conference (Workshops)*. Leuven, Belgium: Association for Computing Machinery, 2013, pp. 170–179. DOI: 10.1145/2460296.2460330. URL: https://doi.org/10.1145/2460296.2460330.

[22] Avraham N. Kluger and Angelo DeNisi. "The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory". In: *Psychological bulletin* 119.2 (1996), p. 254. ISSN: 1939-1455.

[23] Gunther R. Kress and Theo Van Leeuwen. *Multimodal Discourse: The Modes and Media of Contemporary Communication*. An Arnold Publication Series. Arnold, 2001. ISBN: 978-0-340-66292-2. URL: https://books.google.com/books?id=R494tAEACAAJ.

[24] Bibeg Hang Limbu, Halszka Jarodzka, Roland Klemke, and Marcus Specht. "Can you ink while you blink? Assessing mental effort in a sensor-based calligraphy trainer". In: *Sensors* 19.14 (2019), p. 3244.

[25] Kristine Lund. "The importance of gaze and gesture in interactive multimodal explanation". In: *Language Resources and Evaluation* 41.3 (2007), pp. 289–303. ISSN: 1574020X, 15728412. URL: http://www.jstor.org/stable/30204707 (visited on 12/23/2020).

[26] Taylor Martin, Ani Aghababyan, Jay Pfaffman, Jenna Olsen, Stephanie Baker, Philip Janisiewicz, Rachel Phillips, and Carmen Petrick Smith. "Nanogenetic learning analytics: illuminating student learning pathways in an online fraction game". In: *Proceedings of the Third International Conference on Learning Analytics and Knowledge*. Leuven, Belgium: Association for Computing Machinery, 2013, pp. 165–169. DOI: 10.1145/2460296.2460328. URL: https://doi.org/10.1145/2460296.2460328.

[27] Roberto Martinez-Maldonado, Vanessa Echeverria, Doug Elliott, Carmen Axisa, Tamara Power, and Simon Buckingham Shum. "Making the design of CSCL analytics interfaces a co-design process: The case of multimodal teamwork in healthcare". In: *Computer Supported Collaborative Learning 2019*. International Society of the Learning Sciences, 2019, pp. 859–860.

[28] Xavier Ochoa and Federico Dominguez. "Controlled evaluation of a multimodal system to improve oral presentation skills in a real learning setting". In: *British Journal of Educational Technology* 51.5 (2020), pp. 1615–1630. ISSN: 0007-1013. DOI: 10.1111/bjet.12987. URL: https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12987.

[29] Xavier Ochoa, Federico Domínguez, Bruno Guamán, Ricardo Maya, Gabriel Falcones, and Jaime Castells. "The RAP system: automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors". In: *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. Sydney, New South Wales, Australia: Association for Computing Machinery, 2018, pp. 360–364. DOI: 10.1145/3170358.3170406. URL: https://doi.org/10.1145/3170358.3170406.

[30] Xavier Ochoa and Marcelo Worsley. "Augmenting learning analytics with multimodal sensory data". In: *Journal of Learning Analytics* 3.2 (2016), pp. 213–219. ISSN: 1929-7750.

[31] Abelardo Pardo and George Siemens. "Ethical and privacy principles for learning analytics". In: *British Journal of Educational Technology* 45.3 (2014), pp. 438–450. ISSN: 0007-1013. DOI: 10.1111/bjet.12152. URL: https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12152.

[32] Luis P. Prieto, Kshitij Sharma, Pierre Dillenbourg, and María Jesús. "Teaching analytics: Towards automatic extraction of orchestration graphs using wearable sensors". In: *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM, 2016, pp. 148–157. ISBN: 1-4503-4190-X.

[33] Gianluca Romano, Jan Schneider, and Hendrik Drachsler. "Dancing salsa with machines—filling the gap of dancing learning solutions". In: *Sensors* 19.17 (2019), p. 3661.

[34] Stefan Scherer, Marcelo Worsley, and Louis-Philippe Morency. "1st international workshop on multimodal learning analytics: extended abstract". In: *Proceedings of the 14th ACM international conference on Multimodal interaction*. Santa Monica, California, USA: Association for Computing Machinery, 2012, pp. 609–610. DOI: 10.1145/2388676.2388803. URL: https://doi.org/10.1145/2388676.2388803.

[35] Jan Schneider, Dirk Börner, Peter Van Rosmalen, and Marcus Specht. "Presentation trainer, your public speaking multimodal coach". In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. acm, 2015, pp. 539–546. ISBN: 1-4503-3912-3.

[36] Neil Selwyn. "What's the problem with learning analytics?" In: *Journal of Learning Analytics* 6.3 (2019), pp. 11–19. DOI: 10.18608/jla.2019.63.3. URL: https://learning-analytics.info/index.php/JLA/article/view/6386 (visited on 12/23/2020).

[37] Kshitij Sharma, Zacharoula Papamitsiou, and Michail Giannakos. "Building pipelines for educational data using AI and multimodal analytics: A "grey-box" approach". In: *British Journal of Educational Technology* 50.6 (2019), pp. 3004–3031. ISSN: 0007-1013. DOI: 10.1111/bjet.12854. URL: https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12854.

[38] Johannes Wagner, Florian Lingenfelser, Tobias Baur, Ionut Damian, Felix Kistler, and Elisabeth Andr. "The social signal interpretation (SSI) framework: Multimodal signal processing and recognition in real-time". In: *Proceedings of the 21st ACM international conference on Multimedia*. 2502223: ACM, 2013, pp. 831–834. DOI: 10.1145/2502081.2502223.

[39] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. "Convolutional pose machines". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2016, pp. 4724–4732.

[40] Marcelo Worsley. "(Dis) engagement matters: Identifying efficacious learning practices with multimodal learning analytics". In: *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. ACM, 2018, pp. 365–369. ISBN: 1-4503-6400-4.

[41] Marcelo Worsley, Dor Abrahamson, Paulo Blikstein, Shuchi Grover, Bertrand Schneider, and Mike Tissenbaum. "Situating multimodal learning analytics". In: *Proceedings of the Third International Conference on Learning Analytics and Knowledge*. Vol. 2. 2016, pp. 1346–1349. ISBN: 0-9903550-8-X.

[42] Marcelo Worsley and Paulo Blikstein. "Towards the development of multimodal action based assessment". In: *Proceedings of International Conference of the Learning Sciences*. Leuven, Belgium: Association for Computing Machinery, 2013, pp. 94–101. DOI: 10.1145/2460296.2460315. URL: https://doi.org/10.1145/2460296.2460315.