Artificial Intelligence Innovations for Multimodal Learning, Interfaces, and Analytics

Marcelo Worsley

Contents

1 Introduction 20
2 Prior Literature 20
2.1 Multimodal Learning to Support Twenty-First Century Learning Competencies 21
2.2 Multimodal Interfaces to Facilitate Inclusive Learning 21
2.3 Multimodal Analytics to Enable Novel Measures for Learning 22
3 Multicraft 23
3.1 Overview 23
3.2 Multimodal Learning 25
3.3 Multimodal Interfaces 26
3.4 Multimodal Analytics 27
3.5 Summary 27
4 BLINC 28
4.1 Overview 28
4.2 Multimodal Learning 29
4.3 Multimodal Interfaces 29
4.4 Multimodal Analytics 30
4.5 Summary 31
5 Discussion 31
5.1 Multimodal Learning Deserves Multimodal Assessments 31
5.2 Twenty-First Century Skills Benefit from Twenty-First Century Methods 32
5.3 Be Intentional About Keeping Humans in the Loop 32
5.4 Ethical Considerations 33
6 Conclusion 34
References 34

M. Worsley (✉)
Northwestern University, Evanston, IL, USA
e-mail: marcelo.worsley@northwestern.edu

H. Niemi et al. (eds.), *AI in Learning: Designing the Future*,
https://doi.org/10.1007/978-3-031-09687-7_2

1 Introduction

One hallmark of the twenty-first century has been an expansion in the places where meaningful learning takes place. While many discussions of learning had primarily been confined to traditional classrooms and other formal spaces, recent work has reemphasized the important learning that takes place outside of traditional learning settings (Barron and Bell 2015; Pinkard 2019; Vossoughi and Bevan 2014). Some of these spaces involve after-school enrichment programs, open-ended science laboratories, community-based learning experiences, and makerspaces. These spaces can provide learners with authentic and locally situated learning experiences. They can also be used to facilitate learning of a broader set of competencies: critical thinking, collaboration, communication, and creativity, for example. These and other twenty-first century skills have received increased recognition as essential for addressing future societal needs. For example, much research has been conducted to study learner development of twenty-first century skills (Dede 2009), the 4Cs (critical thinking, communication, collaboration, and creativity), and soft skills (Touloumakos 2020). These additional learning contexts and constructs represent important advances in the educational experiences available for today's learners. However, supporting these new types of learning and contexts introduces significant challenges for both learners and educators. Whereas researchers and practitioners have spent decades developing learning experiences and associated measures for competencies like literacy and numeracy, these new contexts and competencies necessitate further research and development. Fortunately, recent advances in the low-cost multimodal sensors can be used to foster new forms of interaction and novel approaches for studying learning that might enable our ability to study, measure, and support these new contexts and competencies.

This chapter will explore the use of multimodal technologies to simultaneously support student learning in nontraditional learning environments and study student learning of these newly emphasized constructs. Two recently developed platforms, Multicraft (Worsley et al. 2021c) and BLINC (Building Literacy in In-Person Collaboration) (Worsley et al. 2021a) will be used to demonstrate how to integrate multimodal interfaces and analytics in K-12 and higher education settings. Each platform supports learners as they practice relatively newly recognized competencies and include a host of multimodal analytics. The two platforms also allow for users to engage in multimodal interactions that utilize speech, eye gaze, tangible blocks, electroencephalography, body pose, and/or facial expressions.

2 Prior Literature

Before moving into a discussion of each platform, this chapter will highlight some pertinent prior research in multimodal learning, multimodal analytics, and multimodal interfaces.

2.1 Multimodal Learning to Support Twenty-First Century Learning Competencies

Within this chapter, we will refer to multimodal learning as being associated with experiences that allow users to (1) engage in learning relevant concepts and ideas through a variety of modalities (e.g., images, videos, text, embodied experiences) and (2) demonstrate their knowledge using a combination of modalities (e.g., speech, written text, drawings, gestures, physical artifacts). The idea of multimodal learning has been a guiding principle within the hands-on, project-based, makerspace, and embodied cognition communities. At the same time, prior research has frequently coupled learning twenty-first century skills, with hands-on, collaborative learning environments that are often supported by computational tools and interfaces. Simply put, many of these contexts emphasize skills of real-world, collaborative problem-solving that are difficult to replicate within a traditional, individual-oriented learning experience. For instance, the process for learning collaboration typically necessitates working in close contact with other individuals and is often situated around a specific unifying real-world problem. Students interact with one another using text, speech, physical artifacts, and gestures, in either colocated or remote settings, for example. Frequently, the means for assessing learning is embedded within the artifact or project that the team creates as opposed to being determined by a written or verbal exam. In summary, attention to learning as multimodal is in alignment with previous calls for epistemological pluralism, equity, accessibility, and inclusion. More generally, researchers have documented the shortcomings of not allowing learners to explore a full set of modalities within a given learning scenario, and the problems with limiting the modalities students are permitted to use to demonstrate their knowledge or learning (Kress 2001; Worsley et al. 2021b).

2.2 Multimodal Interfaces to Facilitate Inclusive Learning

While multimodal learning experiences need not occur through digital technologies, artificial intelligence-enabled multimodal interfaces are becoming an increasingly common strategy for supporting naturalistic interactions between humans and computers (Martinez-Maldonaldo et al. 2017). These interfaces use things like speech-recognition, gesture recognition, and eye tracking, for example, to intelligently interpret the user's intended action. Near the turn of the century, researchers became increasingly intrigued by opportunities to interact with computers using a wide variety of modalities (e.g., speech, eye gaze, gesture, and pen) that typically require some level of artificial intelligence to determine user intent based on an individual modality, or a combination of modalities. Significant decreases in the cost and availability of these multimodal technologies, coupled with the relatively high accuracy of these new tools, fueled considerable advancements in both hardware

and software for capturing and analyzing multimodal data. Developments in video game technology were particularly important contributors to this growth as many computer science researchers explored opportunities to implement multimodal interfaces using the Nintendo Wiimote, Xbox Kinect sensor, and Oculus Rift, for example. The Xbox Kinect sensor included a microphone array for collecting directional audio (to determine who is talking), a depth camera (to estimate object distances), skeletal tracking for up to six individuals (to detect body poses and gestures), and open-source libraries to program the sensors. More recently, researchers have created algorithms that can realize many of those capabilities using a standard web camera, which provides immense opportunities for innovative, low-cost, multimodal interfaces. Researchers and developers create these multimodal interfaces with differing objectives. At times, the interfaces are created to promote accessibility, while in other instances they are developed to enable users to complete their desired tasks more easily. Some common interfaces that feature speech and/or gesture-based input include the smart home technologies available in Amazon Alexa and Google Home, and the touchscreens that are standard within smartphones, tablets, and computers.

2.3 Multimodal Analytics to Enable Novel Measures for Learning

Alongside novel developments in multimodal interfaces, researchers are also developing novel ways to use multimodal data to assess student learning. This specific area of scientific inquiry is called Multimodal Learning Analytics (MMLA) (Blikstein and Worsley 2016; Worsley et al. 2016, 2021b) and refers to ways that multimodal data and computational tools can be employed to model and represent learning within a given environment. The need to study complex learning environments is among the driving motivations for establishing this subfield of learning analytics. Researchers frequently utilize modalities of video, audio, eye gaze and electrodermal activity to look for patterns and forms of interaction that may be hard to identify using traditional learning assessments or through human observation. Additionally, research in MMLA is often concerned with constructs of communication (Ochoa and Dominguez 2020; Ochoa et al. 2018), collaboration (Cukurova et al. 2018; Schneider and Pea 2015; Worsley et al. 2021a), critical thinking (Di Mitri et al. 2020; Oviatt et al. 2015), and creativity (Schneider and Blikstein 2015; Worsley and Blikstein 2018). Across these studies, researchers focus on the combination of audio, gesture, and human-technology interactions to advance theory about collaborative problem solving, communication, creativity and more. MMLA encompasses a broad set of analytic techniques that involve differing levels of human-machine collaboration. In some cases, MMLA analyses involve applying computational techniques to human labelled data. In other cases, researchers might utilize the output from one or more machine learning classifiers to draw inferences

about human learning. In other instances, the analyses may almost exclusively be conducted using machine learning. The unifying perspective across these types of analyses is the realization that multimodal data is essential for supporting the types of inferences that researchers wish to make, and that computational techniques can assist them in providing interpretations of the learning experience.

Prior studies in multimodal learning, multimodal interfaces, and multimodal analytics have individually spurred meaningful contributions to the research community. However, seldom has research from these different areas been integrated with one another. For example, much of the prior work on multimodal learning has tended to rely on traditional measures of student learning. Similarly, work on multimodal interfaces has principally looked at the quality of the user experience, but rarely considered using that same multimodal data to support rich analytics about student learning. Finally, multimodal learning analytics has tended to focus on analyzing data and only seen a select few projects that involve simultaneously using multimodal interfaces together with multimodal analytics. Instead, the multimodal technology has typically only been used to capture data. Intersecting these different areas likely represents the future of learning technologies. This book chapter will describe two examples of tools that sit at the intersection of these three areas. The first, Multicraft, is a multimodal interface for Minecraft that supports collaboration, creativity, computational thinking, and spatial reasoning. The second, BLINC (Building Literacy in In-Person Collaboration) is a platform that uses AI to support real-time collaboration in active learning classrooms, and includes rich, context-specific collaboration analytics. The sections to follow describe each platform in detail and outline their connections to multimodal learning, multimodal interfaces, and multimodal analytics.

3 Multicraft

3.1 Overview

Multicraft is a multiplayer experience for Minecraft that allows for various types of multimodal input. Minecraft is a virtual sandbox game where users can individually or collaboratively design and create buildings, cities, and entire worlds. The platform is sometimes described as a virtual reality space for Legos that has been augmented with some computer programming functionality. Figure 1 includes a picture of a Minecraft world collaboratively created by youth that consists of various puzzles and games. Figure 2 shows a professionally created world that replicates significant portions of Florence, Italy. This particular world aims to allow youth to explore Florence through an interactive virtual reality experience.

Within the current version of the Multicraft platform, users can interact with Minecraft using speech, gestures, eye gaze, tangibles, and even electroencephalography (EEG). The platform was developed to support children with disabilities to

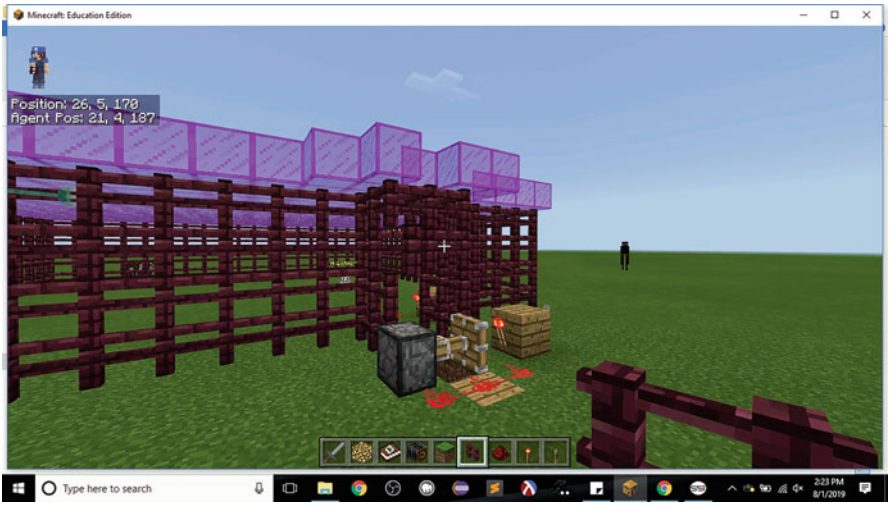

Fig. 1 Picture of Minecraft world created by youth

Fig. 2 Picture of professionally created Minecraft world that replicates Florence, Italy

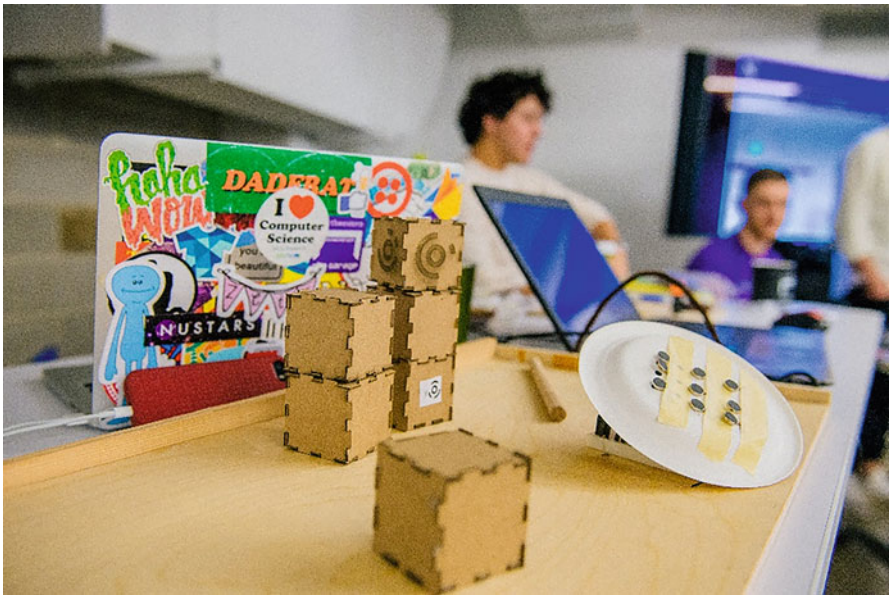

Fig. 3 An early prototype of the tangible interface used within Multicraft

equitably participate in the Minecraft learning experience. Figure 3 shows an early prototype of the tangible interface used within Multicraft (Bar-El et al. 2018).

3.2 Multimodal Learning

As previously noted, Multicraft is designed to be utilized in conjunction with Minecraft, a virtual learning and gaming environment that is popular among youth. The Minecraft learning space allows users to practice several important competencies. Some of these competencies include creativity, problem-solving, spatial reasoning, and computational thinking. Furthermore, it provides the type of virtual world where youth can naturally, and collaboratively, interact with phenomena that connect to any number of disciplines. For example, youth can use Minecraft to create the logic for a computer or use it to create entire cities. Furthermore, the platform is designed to effectively engage and support relative novices, while also being sufficiently generative to allow experts ample opportunities to engage with complex concepts and interactions.

Another hallmark of Minecraft is the opportunity for participants to collaboratively mine, craft, and build within the same virtual world. For example, a group of friends could enter a shared Minecraft world and collectively work on designing a sustainable city over the course of several weeks. Within the game environment, participants are encouraged to communicate with one another through in-game chat,

and control virtual avatars that can interact with one another. Furthermore, educators and computer scientists have developed hundreds of free publicly available lessons that include design challenges, virtual field trips, and more traditional STEM content. These affordances come together to position Minecraft as a learning platform that can advance various twenty-first century competencies.

3.3 Multimodal Interfaces

From a multimodal interface perspective, Minecraft was originally designed to be played with a keyboard and mouse, or a standard gaming controller. In many youth classrooms, it is common to see players use one hand to control the keyboard and the other hand to control the mouse. The Multicraft platform augments the keyboard and mouse-based input, to also include speech, eye gaze, EEG, gestures, and tangibles. Users can select which modalities they wish to employ to complete a given action. An important design principle for Multicraft, however, is to do more than simply replace the existing input modalities using multimodal interfaces. Instead, the platform aims to foster equitable play and leverages computer programming to accelerate some aspects of the gameplay experience. For example, users can say "build a five by ten by eight wood structure here" and Multicraft can utilize a combination of speech recognition, natural language understanding, and eye tracking to instantly build the desired structure where the user is looking. The platform also includes block-based, tangibles input in which a user, or group of users, can manipulate wooden blocks and have their design uploaded to the game in real-time. The tangible block-based input is accomplished using computer vision and relies on a combination of contour detection and color-based tracking. Recent prototypes of the platform also include use of simple hand gestures and EEG. Both approaches are based on machine learning algorithms that can be trained for user-specific gestures or brain activity. The data used to identify hand gestures are from a standard web camera. The EEG data comes from the Muse S headband and includes features from participant brain wave activity. Broadly speaking, Multicraft includes a wide collection of modalities to encourage participants to engage in gameplay using the modalities that best suit them.

These different modalities are important for fostering more equitable and inclusive gameplay and are being researched for their ability to also facilitate improved spatial reasoning and computational thinking. As an example, prior research in spatial reasoning suggests that using spatial language can be a meaningful way to improve spatial reasoning. By encouraging participants to talk to the game using spatial language, we hope to leverage this finding in ways that will result in significant improvements in spatial reasoning. The tangible-based input modality can also confer learning of spatial reasoning. Namely, the use of wooden blocks that exist within the material world, and that are subsequently translated into a 2D representation of the 3D world, can support learners as they practice this process of translating between 2D and 3D representations. Hence, the incorporation of

a multimodal interface can substantively contribute to the goals of multimodal learning of new competencies. Additionally, as we see in the next section, analytics can also help expand how we think about these different competencies and support researchers as they identify and chronicle learner growth with these competencies.

3.4 Multimodal Analytics

The wealth of multimodal data available through Multicraft is also instrumental in supporting analyses of student learning. As an example, this research project includes several hours of data from participants as they engage in Minecraft-focused summer camps and after-school programs. One way for researchers to more tractably navigate human analysis is through the use of computational analyses. Worsley and Bar-El (2019) used log data from the Multicraft server, together with screen recordings of user gameplay, to determine segments in which learners with differing spatial reasoning performance, significantly differed in their in-game interactions. Using this reduced set of data, the authors were able to surface some novel spatial reasoning practices. Worsley and Bar-El describe various ways that students use a combination of explicit and implicit attentional anchors to support the building process within Minecraft. Using eye tracking data, researchers have also highlighted ways that students may practice common spatial reasoning skills within Minecraft, such as perspective-taking and constructing mental representations. At the same time researchers also proposed some spatial reasoning practices that are unique to virtual environments, some of which are based on combinations of well-documented spatial reasoning practices (Andrus et al. 2020). One such practice was error checking, which combines aspects of constructing mental representations and perspective-taking. This project has also used eye tracking data to investigate spatial reasoning practices and identify eye tracking behaviors of learners that exhibit differential performance on common spatial reasoning tasks. Many of these insights are made possible because of the combination of a generative, multimodal learning environment, the utilization of multimodal interfaces, and the computational tools for analyzing data across different modalities.

3.5 Summary

Multicraft is an example of a platform which highlights some of the possibilities for connecting across multimodal learning, multimodal interfaces, and multimodal analytics. Each of these areas is central to the goals and implementation of the platform. Furthermore, the three approaches are integrated to support one another. The next section will present an example designed for the higher education context.

4 BLINC

4.1 Overview

Collaboration is among the most regularly discussed competencies for learners to develop. However, learning institutions seldom offer their students explicit instruction in how to collaborate, or meaningful data around how they are collaborating. A primary goal of the BLINC platform is to provide students with useful insights about how they are collaborating within different contexts. This is achieved by giving users real-time information about how a collaboration is progressing. At a high level, this includes data about how much the group is talking, asking questions, or remaining silent, and the relative distribution of talk among different participants (Fig. 4). The data also includes tracking of user-specified keywords and sentiment classes (Fig. 4). The interface also includes a searchable history of spoken utterances that users can look through for reference. Finally, users can look at discussion content across all groups within the same view and get a summary of verbal contribution frequencies (see Fig. 5).

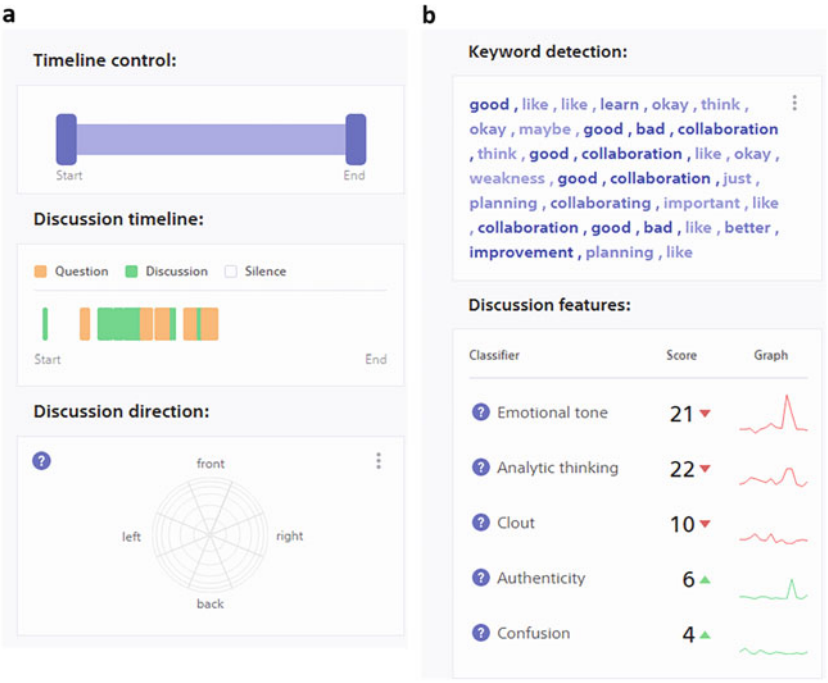

Fig. 4 (**a**) View from BLINC that shows timeline control, portions of questions, discussion, and silence, and the Discussion direction components. (**b**) View from BLINC that shows keyword detection and sentiment analysis

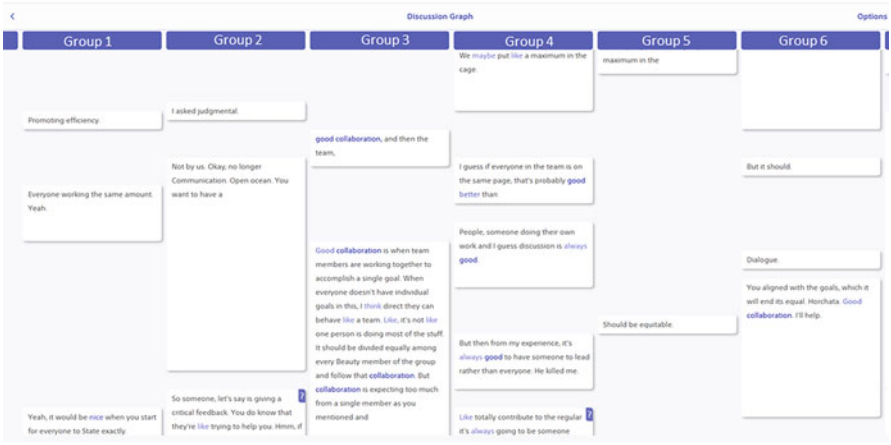

Fig. 5 View from BLINC that shows discussion content for six groups simultaneously

4.2 Multimodal Learning

The BLINC platform was developed amidst growing interest in active learning within institutions of higher education. The term active learning describes a learning environment that contrasts the common practice of learners passively sitting through lectures (Lombardi et al. 2021). Instead, active learning spaces are typified by small group discussions, student-teacher interaction, and limited lecturing. Engaging students in this way can have substantive benefits for student knowledge construction, collaboration, communication, and various other skills that receive significantly less emphasis in traditional lecture-based courses. While this approach is grounded in formative theories from the education research community, instantiating and supporting these types of active learning experiences can present challenges to students and instructors. Instructors may struggle to know how best to support their students within such a format, as it can be difficult to simultaneously have a clear window into all of the small group discussions. At the same time, it can be difficult for learners to get constructive and contextualized feedback from a faculty member who leads a class of more than 50 students. BLINC addresses these challenges through the use of multimodal technologies.

4.3 Multimodal Interfaces

Whereas Multicraft includes a host of multimodal input devices, BLINC primarily uses audio, with an option for video-based input. Users primarily interact with the BLINC system using a web browser which provides them with password-protected access to their current and previous collaboration sessions. Within the

current implementation, audio from collaboration sessions can be captured using two different types of devices. The first is a commercial microphone array called the ReSpeaker Core v2.0. The ReSpeaker includes six microphones to capture audio from up to 5 meters away from the device. The audio capture can be augmented with video from a USB web camera. The BLINC system can accommodate any number of different types of microcomputers through an API that exposes the necessary components for facilitating encrypted data transfer between the microcomputer and the BLINC backend. The second mode for data capture is the microphone from a standard, web-enabled smartphone. Users can access the BLINC webpage and enter a join code for the current discussion. This will subsequently allow them to include their smartphone as one of the audio data collection devices for the group discussion. This feature is particularly salient for higher education contexts where students regularly collaborate outside of class sessions.

In terms of additional interfaces, the platform includes various customizable visualizations and data representations that can support participant sensemaking around their data. The specific time ranges can be adjusted using a slider, and nearly all of the visualizations provide drill down capabilities that take the user to the underlying text associated with a given data point or data segment.

4.4 Multimodal Analytics

The various capabilities offered through the BLINC platform are heavily dependent on multimodal analytics. Even though most of the data being analyzed comes through a single modality (i.e., audio), computational tools and techniques allow for that data to be transformed into several meaningful data points. This section will outline some of those capabilities.

The analytic pipeline begins with the collection of multichannel audio. Each of the six microphones captures audio from the surrounding area. That multichannel audio is used to compute the direction of arrival based on differences in the amount of time it took for a given utterance to reach each of the different microphones. The audio data subsequently undergoes speech recognition. Speech recognition translates from audio into text. The text is later used for various text processing tasks. BLINC also includes speaker diarization. Speaker diarization provides an estimation of who said each utterance. The utterances are labelled with generic titles (e.g., Speaker 1, Speaker 2, etc.). While the platform can support direction of arrival to an accuracy of 20–30 degrees, speaker diarization offers an important augmentation in settings where participants are not stationary, and when users are collecting data through their smartphones. The results from speech recognition also include timestamps on a per utterance basis, and estimated punctuation. Both pieces of information are useful in quantifying the distribution of talk among different team members and the relative timing and distribution of silence, questions, and discussion. As previously noted, the primary output from speech recognition is an estimated transcript of what group participants said. That transcript is used to

support keyword detection. For example, in a class on educational technology, an instructor could specify a collection of keywords: creativity, innovation, technology, ethics, and data. The system would annotate each utterance containing one of those keywords and keep a count of each keyword that appears in the transcript. Furthermore, the system has integrated topic modeling (McCallum 2002). Users can, provide a custom set of documents to train a course- or context-specific topic model and subsequently use that model to examine and chronicle how much group discussion aligns with the different topics. It can also represent how groups are transitioning between the different topics.

4.5 Summary

The BLINC platform sits on top of several computational techniques for analyzing and extracting meaning from audio. While audio is the primary modality, the platform finds several ways to deconstruct that data into useful insights for learners and educators. In so doing, the platform fills an important practical gap of supporting active learning in large enrollment classes and allowing users to explore their collaboration literacy outside of the classroom. Hence, the platform aims to bring together the need for collaborative, active learning, the challenge of facilitating such learning, and the opportunities for utilizing multimodal data and analytics in ways that can support researchers, learners, and educators.

5 Discussion

Multicraft and BLINC provide a glimpse of potential innovations that integrate multimodal learning, interfaces, and analytics. Each platform provides tangible benefits for both users and researchers. At the same time, the pair of projects also highlight a few commonalities that are described in the subsequent sections.

5.1 Multimodal Learning Deserves Multimodal Assessments

The design of Multicraft and BLINC are both informed by the realities of new types of learning experiences. BLINC is designed to support collaborative learning environments where students are actively engaged in discussions with their peers and the course instructors. BLINC also supports student collaboration in out-of-school contexts, through the "bring your own device" (BYOD) feature. Both features speak to the idea of students engaging in what we are loosely calling multimodal learning. Similarly, Multicraft, or Minecraft more broadly, is a virtual learning environment where players can collaboratively engage in hours of creative

designing, mining, crafting, and exploring. While researchers have looked at these types of learning environments through traditional assessments and constructs, those constructs fail to do justice to the types of learning and competencies that the spaces make available. Furthermore, asking students to learn and practice material through a variety of modalities, and subsequently restricting assessments to a single modality represents a contradiction to the design and motivation of multimodal learning experiences.

5.2 *Twenty-First Century Skills Benefit from Twenty-First Century Methods*

Some of the competencies supported through BLINC and Multicraft include collaboration, communication, spatial reasoning, and computational thinking. Researchers have explored various methods for studying these, with many relying on traditional techniques from quantitative and qualitative research traditions. These have been beneficial in furthering our understanding of these constructs, but part of what we see with these two platforms is the need for novel methods for examining these different skills. For Multicraft, while we could administer a typical mental rotation test, such a test becomes highly decontextualized and lacks authenticity and contextual validity. Instead, leveraging computational techniques from eye-tracking data, for instance, can surface the visual spatial anchors that participants may use as part of the building process. Similarly, EEG data might highlight aspects of student concentration and focus that go undetected using most traditional tests and analytic approaches. In the case of BLINC, the platform can support temporal and group-level inferencing about how a group is collaborating. This goes well beyond what one might get from simply having participants complete pre- and post-tests about their collaboration preferences, for example.

5.3 *Be Intentional About Keeping Humans in the Loop*

A final unifying idea to discuss with regard to Multicraft and BLINC is their intentionality in keeping humans in the loop. Many discussions of artificial intelligence gravitate towards fully automated systems that seemingly replicate human reasoning. Neither Multicraft nor BLINC follow this paradigm. Instead, the platforms reflect inclusion of human decision-making and inference throughout their design and use. They are also intentional about avoiding explicit prescriptions or labelling of individuals and make an effort to present data in context. Many of these approaches are most readily apparent in BLINC. First, the BLINC platform includes considerable customization that can cater the data representations to the specific keywords that the students or instructor wish to focus on, for example. BLINC also avoids generating prescriptions or recommendations around an ideal collaboration

style. For instance, the data representations concerning verbal contributions do not include suggested target values. Instead, instructors and participants are encouraged to use the data in conjunction with their knowledge of the specific learning context and group. This combination of information can help them reflect upon and modify their collaboration practices. Additionally, the ability to drill down into the specific utterances that underlie the visualizations means that humans have an opportunity to interrogate the representations and determine which pieces of data necessitate significant user action. In these ways, these systems aim to simultaneously take advantage of the power of artificial intelligence and the complex reasoning patterns that humans exhibit. Certainly, as society moves into scenarios where people are practicing and evaluating new competencies, it will be beneficial to leverage both of these forms of intelligence, or as Doug Engelbart would say, to "co-evolve" human-computer intelligent systems.

5.4 Ethical Considerations

As society continues to explore the various innovations that might be had through integrating multimodal learning, interfaces, and analytics, it is important to touch on some ethical considerations that can be used to protect participants. Worsley, Martinez-Maldonado, and D'Angelo (Worsley et al. 2021b) include a detailed discussion of 12 core MMLA commitments that span the research pipeline. Their discussion outlines commitments related to data collection, data analysis, and data dissemination. Most salient under the idea of data collection is being circumspect and transparent about what multimodal data is being collected and providing ways for participants to control when that data is being collected. Within the data analysis portion, two commitments that stand out are related to thorough, consistent, and transparent data modeling, and creating opportunities for participants to provide feedback and reflection within the data analysis process. Broadly speaking these two commitments aim to minimize researcher or algorithmic bias. Finally, with regard to dissemination, the authors argue for researchers to develop multimodal systems that provide tangible benefits to research participants. This commitment is not intended to undercut the overall value of research, but to instead advocate for researchers to embark on studies that can potentially confer meaningful benefits to participants, whenever possible. Researchers and designers of multimodal systems should elevate the needs of users. Moreover, the field must carefully consider how this work might feasibly be integrated into ecological settings and how it might scale from classrooms, to schools, to entire districts. These points of integration cannot merely be about the technologies, but must also center ethics.

6 Conclusion

Artificial Intelligence is quickly becoming an integral part of our lived experiences. From speech recognition to computer vision and natural language processing, AI is poised to make a significant impact on the future of learning. One particularly impactful point of integration could be in bridging among multimodal learning, multimodal interfaces, and multimodal analytics. This chapter explored some examples that effectively merge these three areas in ways that support student learning of novel competencies. Notwithstanding, this chapter suggests that truly fomenting student growth in these newly dubbed competencies may require expanding the modalities and analytic techniques that researchers employ.

References

Andrus, B., Bar-el, D., Msall, C., Uttal, D., & Worsley, M. (2020). Minecraft as a Generative Platform for Analyzing and Practicing Spatial Reasoning. *Spatial Cognition XII*.

Bar-El, D., Davison, L., Large, T., & Worsley, M. (2018). Tangicraft: A Multimodal Interface for Minecraft. *20th International ACM SIGACCESS Conference on Computers and Accessibility*.

Barron, B., & Bell, P. (2015). Learning environments in and out of school. *Handbook of Educational Psychology*, 323–336.

Blikstein, P., & Worsley, M. (2016). Multimodal Learning Analytics and Education Data Mining: using computational technologies to measure complex learning tasks, *3*(2), 220–238. https://doi.org/10.18608/jla.2016.32.11

Cukurova, M., Luckin, R., Millán, E., & Mavrikis, M. (2018). The NISPI framework: Analysing collaborative problem-solving from students' physical interactions. *Computers & Education*, *116*, 93–109. https://doi.org/10.1016/j.compedu.2017.08.007

Dede, C. (2009). Comparing Frameworks for "21st Century Skills," 1–16.

Di Mitri, D., Schneider, J., Trebing, K., Sopka, S., Specht, M., & Drachsler, H. (2020). Real-Time Multimodal Feedback with the CPR Tutor. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education* (pp. 141–152). Cham: Springer International Publishing.

Kress, G. (2001). Multimodal teaching and learning: The rhetorics of the science classroom. Retrieved from https://doi.org/10.1002/9781405198431.wbeal0815/full

Lombardi, D., Shipley, T. F., Bailey, J. M., Bretones, P. S., Prather, E. E., Ballen, C. J., ... Docktor, J. L. (2021). The Curious Construct of Active Learning. *Psychological Science in the Public Interest*https://doi.org/10.1177/1529100620973974

Martinez-Maldonaldo, R., Buckingham-Shum, S., Schneider, B., Charleer, S., Klerkx, J., & Duval, E. (2017). Learning Analytics for Natural User Interfaces. *Journal of Learning Analytics*, *4*(1), 24–57. https://doi.org/10.18608/jla.2017.41.4

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. Retrieved from http://mallet.cs.umass.edu

Ochoa, X., & Dominguez, F. (2020). Controlled evaluation of a multimodal system to improve oral presentation skills in a real learning setting. *British Journal of Educational Technology*.

Ochoa, X., Dominguez, F., Guamán, B., Maya, R., Falcones, G., & Castells, J. (2018). 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* (pp. 360–364). New York, NY, USA: ACM. https://doi.org/10.1145/3170358.3170406

Oviatt, S., Hang, K., Zhou, J., & Chen, F. (2015). Spoken interruptions signal productive problem solving and domain expertise in mathematics. *ICMI 2015 – Proceedings of the 2015*

ACM International Conference on Multimodal Interaction, 311–318. https://doi.org/10.1145/2818346.2820743

Pinkard, N. (2019). Freedom of movement: Defining, researching, and designing the components of a healthy learning ecosystem. *Human Development*https://doi.org/10.1159/000496075

Schneider, B., & Blikstein, P. (2015). Unraveling students' interaction around a tangible interface using multimodal learning analytics. *Journal of Educational Data Mining*, *7*(3), 89–116. Retrieved from https://jedm.educationaldatamining.org/index.php/JEDM/article/view/JEDM102

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

Touloumakos, A. K. (2020). Expanded Yet Restricted: A Mini Review of the Soft Skills Literature. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2020.02207

Vossoughi, S., & Bevan, B. (2014). *Making and tinkering: A review of the literature*. *National Research Council Committee on Out of School Time STEM*. Washington, DC, DC.

Worsley, M., Abrahamson, D., Blikstein, P., Grover, S., Schneider, B., & Tissenbaum, M. (2016). Situating Multimodal Learning Analytics. *International Conference for the Learning Sciences 2016*, *2*, 1346–1349.

Worsley, M., Anderson, K., Melo, N., & Jang, J. (2021a). Designing Analytics for Collaboration Literacy and Student Empowerment. *Journal of Learning Analytics*. Retrieved from https://northwestern.box.com/s/1ssb45coayr4h76e09gxk6ktj1qhka6p

Worsley, M., & Blikstein, P. (2018). A multimodal analysis of making. *International Journal of Artificial Intelligence in Education*, *28*(3), 385–419. https://doi.org/10.1007/s40593-017-0160-1

Worsley, M., Martinez-Maldonado, R., & D'Angelo, C. (2021b). A New Era in Multimodal Learning Analytics: Twelve Core Commitments to Ground and Grow MMLA. *Journal of Learning Analytics*, 1–18. https://doi.org/10.18608/jla.2021.7361

Worsley, M., Mendoza, K., Mwiti, T., Zhen, M., & Jiang, M. (2021c). Multicraft: A multimodal interface for supporting and studying learning in Minecraft. In *2021 Human Computer Interaction International*.