



Multimodal analysis of interaction data from embodied education technologies

Candace Walkington¹  · Mitchell J. Nathan² · Wen Huang¹ · Jonathan Hunnicutt¹ · Julianna Washington¹

Accepted: 4 June 2023
© The Author(s) 2023

Abstract

The emergence of immersive digital technologies, such as shared augmented reality (shAR), virtual reality (VR) and motion capture (MC) offers promising new opportunities to advance our understanding of human cognition and design innovative technology-enhanced learning experiences. Theoretical frameworks for embodied and extended cognition can guide novel ways in which learning in these environments can be understood and analyzed. This conceptual paper explores a research method in Educational Technology—multimodal analysis for embodied technologies—and provides examples from shAR, VR, and MC projects that use this approach. This analysis involves tracking learners’ gestures, actions on physical and virtual objects, whole body movements and positions, and their talk moves, in addition to other relevant modalities (e.g., written inscriptions), over time and across space. We show how this analysis allows for new considerations to arise relating to the design of educational technology to promote collaboration, to more fully capture students’ knowledge, and to understand and leverage the perspectives of learners.

Keywords Augmented reality · Virtual reality · Motion capture · Embodied cognition · Multimodal analysis · Gesture

Introduction

The emergence of immersive digital technologies, such as augmented reality (AR), virtual reality (VR) and motion capture (MC) offers promising new opportunities to advance our understanding of human cognition and design innovative learning experiences. Indeed, researchers have argued that these technologies “have the potential to do for gestures what writing did for speaking” and will “transform how people generate, disseminate, and interact with knowledge” (Dimmel & Bock, 2019, p. 2). AR platforms now support experiences that enable people to directly physically and perceptually engage with shared objects. We refer to these new forms of AR technology as *shared holographic AR* (shAR). shAR

✉ Candace Walkington
cwalkington@smu.edu

¹ Department of Teaching & Learning, Southern Methodist University, Dallas, USA

² Department of Educational Psychology, University of Wisconsin–Madison, Madison, USA

enables multiple learners to view, manipulate, and reason about *the same* objects projected as holograms in a joint three-dimensional collaborative space, based on each person's visual perspective, using intuitive hand gestures. Shared VR technologies are somewhat similar, but they allow learners to collaborate around shared objects rendered in a fully virtual space. Both AR and VR technologies can offer unique affordances for distance learning, as students do not need to be co-located. Finally, MC technologies not only allow for the hand tracking that powers the gesture-based interfaces of contemporary AR and VR, but can direct learners' body motions, and then relate these body motions to particular concepts or learning standards.

There are important similarities and differences between these three technologies. AR is a semi-immersive technology where a computer-generated layer is placed on an individual's view of the real world (Blair, 2016). The layer is semi-transparent, allowing individuals to simultaneously manipulate virtual objects and see the real world around them (Blair, 2016). VR is a fully immersive, three-dimensional, computer-generated environment that simulates real or imaged situations that individuals can interact with (Aukstakalnis & Blatner, 1992; Milgram & Kishino, 1994; Onyesolu & Eze, 2011). Collaborators often appear as computer-generated avatars in VR. Motion capture digitizes the motion of a person's different moving parts from a live performance (Menache, 2011), which allows the learning system to provide immediate feedback based on the individual's tracked body movements (Johnson-Glenberg et al., 2014). All three technologies support embodied learning beyond traditional instructional approaches. The most significant difference is the relationship between learners and the learning environment. Learners may show different interaction and reasoning patterns with these different technologies, depending on the level of immersion, how collaborative interactions are structured, and how physical movements are used and recorded.

This conceptual paper explores a research method for analyzing data from educational technology research called *multimodal analysis* (Jewitt, 2017). We describe a particular approach within this umbrella method, where multimodal analysis is adapted for the context of embodied, interactive educational technology systems where there are shared interactions between learners. This differs from how one might analyze data from immersive technologies that are not collaborative—where learners play games or experience simulated environment without sustained embodied interactions with other humans (collaborators or instructors). By “embodied interaction” we highlight that interlocutors must have access to interactional information beyond others' speech or writing—such as gestures and movements. We refer to this approach *multimodal analysis for embodied technologies* (MAET) and provide examples from shAR, VR, and MC projects that use this approach. Our primary aim is to show several ways that MAET allows for deep analyses of the embodied and extended nature of cognition—or the ways in which learners use their bodies to work together and establish shared meaning in technology-enhanced environments.

Theoretical framework

The learning potential for emerging interactive technologies like shAR, VR, and MC draws on the framework of *embodied cognition* (Wilson, 2002)—the idea that all thoughts are grounded in perception, action (including epistemic and pragmatic actions and gestures), and spatial systems. Embodied cognition has risen as a particularly powerful framework to describe learning in mathematics education. For example, research shows that multisensory perceptions can

help learners perceive structures in math that might not be available from symbolic representations (Gerofsky, 2007; Sinclair, 2005) and that students' math knowledge can be revealed through their gestures (Edwards, 2009; Kim et al., 2011; Ng & Sinclair, 2015a, 2015b; Pier et al., 2019), action-oriented language (Nathan et al., 2014), and body-based and spatial metaphors (Lakoff & Núñez, 2000; Roth, 2011). Some members of a branch of embodied cognition called *enactivism* (Varela et al., 1991/2017) make the bold statement that "learning *is* moving in new ways" (Abrahamson & Sánchez-García, 2016, p. 233). This means that when students perform new embodied behaviors, they are engaging *concomitantly* with movements *and* emergent cognitive structures that enable perceptual processes, which guide the motor enactment of these movements.

Increasingly, enactivist scholars have attended to the collaborative nature of embodied behaviors, especially as they relate to joint attention (Shvarts & Abrahamson, 2019). Dual eye tracking can reveal the formation of intersubjectivity among teacher-student dyads where joint (visual) attention, along with strategic pedagogical prompts, enables the emergence of joint action and joint attention. This in turn signals a unified, dynamical coordination between the teacher's and student's attention and actions.

However, research on embodied learning has not broadly embraced the collaborative nature of learning environments (see Abrahamson et al., 2020), which has become increasingly important as technologies like shAR allow multiple learners to simultaneously engage. Research on *extended and distributed cognition* (Clark & Chalmers, 1998; Salomon, 1993) reveals some of the ways in which knowledge and learning can be understood as stretched over people, places, and objects, rather than describing its location as merely in the head of individuals. For example, observations of complex, multiday, project-based engineering classes reveal how teachers exercise numerous mechanisms for fostering and maintaining cohesion. Here, cohesion refers to the integrated understanding students' have of key STEM concepts in the project. These concepts are presented in ways that vary across time, spaces, and artifacts as a single scientific concept is manifest in various symbols, drawings, software simulations, and material forms (Nathan et al., 2017; Walkington et al., 2014). Detailed analyses of these teaching and learning interactions highlight how key STEM concepts are embedded in these multimodal contexts and how knowledge can be distributed across a multitude of physical, digital, cultural, and social resources. This lays the groundwork for an embodied theory of transfer, where, through careful coordination, "both teachers and learners engage embodied processes as they map invariant relations across various modal forms... to apply prior modes of perceiving and acting to new contexts and to create movements that will activate those invariant relations through transduction" (Nathan & Alibali, 2021, p. 50). Describing cognition as extended and distributed is particularly relevant when considering how people collaborate with each other when supported by digital tools. In addition, digital tools have become increasingly well-suited for supporting the distributed and embodied nature of cognition as hand/skeleton tracking, eye tracking, touch screens, and physiological sensors have become more powerful, accurate, and affordable, and tools for managing and analyzing the large amounts of data have become more efficient and user-friendly.

Multimodal analysis for embodied technologies

The increasingly embodied and distributed affordances of educational technology has created a need for methods of analysis to examine the rich interactions learners have in these contexts. Multimodal analysis (Alibali & Nathan, 2012; Jewitt, 2017; McNeill, 1992; see

also Nathan et al., 2017) is a research method originally drawn primarily from the literature on gesture. The method we detail, MAET, is adapted from multimodal analysis to take into account other embodied, distributed, and action-based ways in which students engage with mathematics learning materials and technologies. A *modality* is a channel that people used to communicate—like speech, writing, gesture, etc.—and multimodal analysis attempts to account for the varied modalities that arise as collaborative problem-solving is carried out within or outside of school. In MAET, we seek to do away with the tendency to privilege particular modalities that have been traditionally emphasized in academic settings—like speech and written work—and instead capture a broader concept of the many forms of embodied knowledge that learners express.

MAET first involves a close examination of learners' *gestures*, which we define as spontaneous or planned movements of the hands or arms that often accompany speech and that sometimes convey spatial or relational information. Gestures have been the subject of both observational and intervention research because of their relationship with thinking, social cuing, and cognitive development (Goldin-Meadow, 2005). Several important types of gestures have been identified, including *pointing gestures* that indicate positional information (e.g., pointing to a triangle on the board), *depictive or iconic gestures* which form shapes or show movements using the hands (e.g., forming a triangle with two palms and thumbs), and *beat gestures* which emphasize ideas presented in speech (e.g., a vertical downward movement of the hand to emphasize the word *congruent*; McNeill, 1992; Alibali & Nathan, 2012). A review by Alibali and Nathan (2012) found converging evidence that representational gestures, which depict objects and processes, exhibit mental simulations of actions, perception, and conceptual metaphors.

There are many existing multimodal data analysis approaches that have been proposed by researchers (see Jewitt et al., 2016). Although most of these approaches can be used to handle actions or speech in collaboration analytics, different approaches have different aims and draw upon different theories of learning and interaction. For example, the social semiotics approach (Bezemer & Kress, 2015; Jewitt & Henriksen, 2016) recognizes the relationship between social actors and the agency of their communication. In this approach, purposive interaction establishes actors' relationships. The conversation analysis approach (Deppermann, 2013; Mondada, 2019) explores how participants organize their actions sequentially. The base assumption of this approach is that social acts can be understood by the acts before and after them. The multimodal (inter)action analysis approach (Norris, 2016) investigates how various semiotic resources are introduced into and make up social interaction, identities, and relationships, and this approach emphasizes interactive actions are mediated by space and artefacts. The primary affordance of MAET, however, is that it is aligned with theories of embodied cognition, which have enormous potential for designing immersive learning environments using interactive and movement-based technologies. Other multimodal data analysis methods do not have this focus. MAET is particularly useful for research that relates to using physical or virtual manipulations or actions to help people learn. Although other approaches can include actions, the value of actions is usually related to facilitating communication (e.g., Bezemer & Kress, 2015; Mondada, 2019; Norris, 2016). While communication is important to learning, embodied approaches highlight that action itself also can help students learn. This valuing of action as a conceptual tool is not as accounted for in other multimodal traditions.

An analysis of these categories of gestural interactions has been traditional in multimodal analysis, however recent work has identified further important classes of gesture used during collaboration in technology-enhanced environments. One particularly important form of depictive gestures, *dynamic depictive gestures*, show a motion-based

transformation of an object through multiple states. For example, a student might make a rectangle with their thumbs and forefingers, and then slant it to one side to show a triangle being taken off one side and added to another. Strong associations have been found between dynamic gestures and valid mathematical and spatial reasoning (Göksun et al., 2013; Nathan et al., 2014; Newcombe & Shipley, 2012; Pier et al., 2019; Uttal et al., 2013).

In addition, gesture studies offer an important link between individualized and social forms of embodiment. This is because, while gesture production has well-established cognitive benefits for the individual actor (e.g., Goldin-Meadow, 2005), gesture production is facilitated when speakers operate in a social context (e.g., Goodwin, 2000; Moll & Tomasello, 2007; Vygotsky, 1978), even when the speakers cannot see one another (Alibali et al., 2001). An important class of gestures borne from multi-person interactions are *collaborative gestures*—i.e., gestures that are directly related to or that build upon the gestures of interactional partner(s) (Walkington et al., 2019a, 2019b). Observations of collaborative learning reveal that learners regularly repeat each other's gestures through echoing (when one student repeats a representational gesture of another student) or mirroring (when two or more students simultaneously make similar gestures) gestures. Learners also build on one another's gestures through alternation gestures that respond to other gestures, and learners physically co-represent a single object using joint gestures. Several studies suggest that gesturing collaboratively can be associated with higher performance than making individual gestures (Vest et al., 2020). An examination of collaborative gestures differs from examining individual gestures as the prior interactional context of gestures during reasoning episodes is explicitly taken into account.

MAET also takes into account body position, gaze, and other body language, as well as *physical movements* around a learning environment, to capture learners' joint focus of attention and opportunities for engagement. For example, by tracking learners' gaze we can see how their attention shifts between their collaborator's faces, gestures, and the virtual or physical objects they are working with. By paying attention to body position, we can understand the perspective the learner is taking on the scene or simulation they are interacting with, as well as how they are positioned or not positioned for collaboration with others due to physical proximity. Students can become immersed in mathematical representations and use body positions to see the object from different perspectives (Dimmel et al., 2021). A collaborative view of physical movements differs from an individual view of such movements in that we are concerned with how bodies are positioned with respect to other bodies (rather than only with respect to instructional objects), how learners consider issues of perspective in a collaborative situation, and how body cues are used to focus joint attention.

Further, MAET examines how these actions are coordinated with speech, with a particular emphasis on *collaborative talk moves* (Andrews-Todd et al., 2019). During collaboration, gestures operate synchronously with speech and other movements (e.g., lifting a pen), acting as a mechanism to create cohesion and bind conversational elements (Enyedy, 2005; Koschmann & LaBarron, 2002). When using collaborative talk moves in technology-enhanced environments, learners negotiate, plan, represent, and exchange ideas using language. These discourse moves are related to profiles of collaboration, which in turn predict task performance measures (Andrews-Todd & Forsyth, 2020). Collaborative talk moves differ from individual talk moves in that they take into account the interpersonal and multimodal context in which utterances are made, they attend to how people respond to and build off of one another's multimodal reasoning, and account for the kinds of historical multimodal interactions preceded each talk move.

While collaborative talk moves have been important in our work, there are many other ways to code learners' verbal language. For example, we have found "operational speech" where learners express themselves performing operations on mathematical objects through spoken language, to be an important category in our multimodal analyses (see Pier et al., 2019). Scholars have also utilized text mining tools like LIWC (Pennebaker et al., 2007) and Coh-Metrix (McNamara et al., 2013) to automatically code the textual properties of students' speech—especially their use of pronouns, verbs, situation model construction, and logic as they express and explain mathematical ideas (e.g., Nathan et al., 2020).

Finally, MAET involves looking carefully at learners' *actions with objects and tools* in their environment, including virtual objects like holograms. Actions on objects are defined here as physical manipulations of objects like turning a plastic cylinder on its side or clicking to rotate a triangle in a Dynamic Geometry System (DGS). They are distinct from gestures in that the hands are used as a vehicle to manipulate physical or virtual objects and are pragmatic actions, rather than having the hand gestures themselves represent or index these objects. In a DGSs, learners can use menu commands to construct line segments within a geometric figure, or learners can orient a manipulative towards a particular person. These would not be considered gestures, but such actions may have an important embodied influence on learning (Goldin-Meadow & Beilock, 2010). Collaborative actions on objects could be cases where a student repeats or attempts to anticipate an action an interactional partner is going to take (e.g., resizing a virtual cube as their partner mentions resizing in speech), and these may also occur differentially depending on the collaborative situation. A collaborative view of manipulating objects and using digital tools differs from an individual view by exploring how these actions are preceded by previous multimodal actions of collaborators, and how current actions are performed in particular, embodied ways because collaborators are present.

These four interactional categories are often captured through video recordings, screen recordings, and motion capture technologies. Video cameras from different perspectives are aimed on multiple students working together and are synced with first-person camera feeds from each students' device (e.g., a headset) when appropriate. Body and hand positioning coordinates can be captured from motion capture systems, as we do in our work on *The Hidden Village*. In addition, for embodied environments that use a screen (rather than a headset), the screen itself is recorded with the actions students take through the screen in the environment. Once the different video feeds have been synced together, analysis can begin. We also sometimes utilize log files collected by the technology system in conjunction with video-based data streams.

These four multimodal streams of data (gestures, movements, speech, actions) are coded first at coarser grain sizes across a corpus, after video recordings are divided up into "episodes" that represent student exploration of one task or idea. The codes from these episodes can be quantitative in nature—for example, dichotomous codes such as whether a collaborative talk move was present (coded as 0/1), and frequencies, such as how many gestures were produced, etc. In this way quantitative patterns can be determined from the data using predictive and correlational analyses (e.g., Walkington et al., 2019b, 2022). Analyses can also be done on how these individual classes of multimodal moves are used either in sequence (e.g., a collaborative structure where students perform actions on objects first, and then discuss their ideas verbally) or simultaneously (e.g., pointing to a virtual object while a collaborator is talking). For these analyses, episodes may be broken down further into sub-episodes that coalesce around learners exploring one idea or concept. We have used qualitative video analysis software like Transana, NVivo, and V-Note to perform these analyses, sometimes followed by quantitative analyses in software like R or SPSS.

After this broader analysis, particular video instances are “zoomed in” on, creating multimodal transcripts of key trouble spots, learning opportunities, and collaborative exchanges (see Goodwin, 2000; Sacks et al., 1974). This is especially important when the volume, sequencing, and complexity of the multimodal exchanges in an individual episode is not well-captured by broader codes. When we create these transcripts, we often use overlaid lines or arrows to highlight gestures, actions, movements, and eye gaze, so that attention is focused on these other modalities for communication, and speech is presented as just another channel. We also explicitly transcribe the knowledge captured in gesture and movement by describing the gestural forms and types of movement in words.

An important question is how MAET is different from methods used to analyze “unplugged” embodied learning, or from typical multimodal analyses. First, MAET capitalizes on the affordance of being able to dynamically manipulate virtual objects in real-time, with precise and instantaneous measurement of key values. These technologies also allow for scale to be changed in new and exciting ways. The coding categories for MAET were formulated with these particular affordances of embodied technologies in mind. MAET also capitalizes on the increasingly gestural-based interfaces of embodied technologies, and their important role in mathematical reasoning and in collaboration. Further, MAET’s focus on body positioning and viewpoint becomes increasingly important in virtual environments like VR where learners must purposefully navigate and achieve collaborative discourse with their virtual rather than physical bodies. It is also important as technologies can allow students to instantly take different perspectives (e.g., teleport to an area in a virtual world or instant switch from a view where a coordinate plane is position vertically as a blackboard would be or horizontally atop the floor). Finally, MAET leverages the coordination of many different data streams (distant camera, first person device or screen-recording, log and position tracking data), which have not been traditionally available in unplugged embodied learning.

Our research has used MAET to explore learning in three contexts—motion capture, shared VR, and shAR. We will discuss each of these projects as cases of the ways in which MAET can be used to explore data from learners engaging with educational technology.

MAET in a motion capture game

We engaged in a MC project where high school students were engaged in a video game environment called *The Hidden Village*. In the game, which utilized the visual novel genre, the players crash-land on an alien planet, and are asked to perform different arm motions as they meet different villagers (Fig. 1). Their motions are captured by a Kinect™ camera (although a current version of THV uses a laptop camera) to determine if players make the desired poses. Each set of arm motions is related to a geometry conjecture that students would subsequently have to prove as being always true or ever false (Fig. 2). For example, the students might make arm motions where their arms make a growing triangle (to show mathematically *similar* triangles) before being asked to prove or disprove the conjecture “Given that you know the measure of all three angles of a triangle, there is only one unique triangle that can be formed with these three angle measurements.” This conjecture is false, as triangles with different sizes can have the same angle measurements (i.e., similar triangles). The game also allows learners to create their own arm motions and program them into the game to correspond to new geometric ideas, using the Conjecture Builder feature (Fig. 3; see Walkington et al., under review).

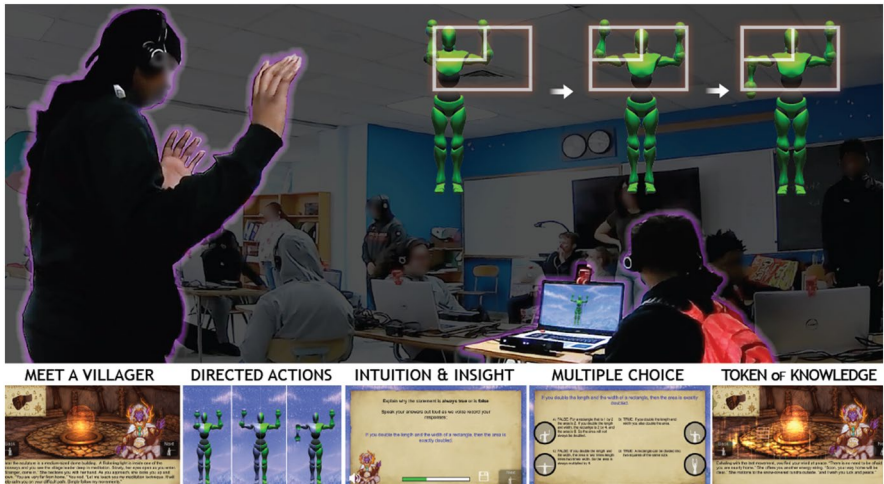


Fig. 1 Game flow: meet villager, perform *directed actions*, followed by free-responses to the given conjecture and selecting a multiple-choice response, and receiving a token of knowledge

Conjecture	Relevant Actions	Intended Relevance of Actions
1. Given that you know the measure of all three angles of a triangle, there is only one unique triangle that can be formed with these three angle measurements. (False)		Poses show a triangle getting larger.
2. The area of a parallelogram is the same as the area of a rectangle with the same base and width.		Poses show a rectangle, and then a parallelogram where area is re-organized.
3. The diagonals of a rectangle always have the same length.		Poses show two right triangles that make up the diagonals of a rectangle.
4. The opposite angle of two lines that crosses are always the same.		Poses show sets of vertical angles that are forced to be equal no matter how two lines are positioned.

Fig. 2 Examples of relevant action sequences programmed into The Hidden Village game

This allows the *students* to determine how different physical motions might correspond to geometry ideas.

In Fig. 4, we present an example of MAET of learners using the Conjecture Builder to create arm motions to accompany a geometry conjecture about quadrilaterals. We approached this analysis by examining how the multimodal resources that students used to reason about their focal game conjecture—which here was “The diagonals of a rhombus bisect the angles at all four vertices” changed over time. The video footage of students working with the Conjecture Creator was segmented into different episodes as they responded to different portions of the creation task (e.g., explore the focal conjecture, come up with a formal proof for the focal conjecture, come up with poses for the focal conjecture, etc.). Figure 4 shows the MAET of the episodes where the two learners were coming up with arm poses to accompany the conjecture. This episode was coded with different types of collaborative gestures, such as in Lines 1–4 when the students are mirroring each other’s gestures. It was also coded with collaborative talk moves, like Representing talk moves where learners build representations of the problem and formulating hypotheses (Lines 3–7) and Negotiating talk moves that express agreement or disagreement (Line 8). The MAET of this episode shows how much of the conversation is expressed in rich gesture sequences rather than in precise mathematical speech. We coded this theme as “gesturally embodied argumentation spurred by pose creation,” and examined its conditions of incidence across the corpus (see Walkington et al., under review).

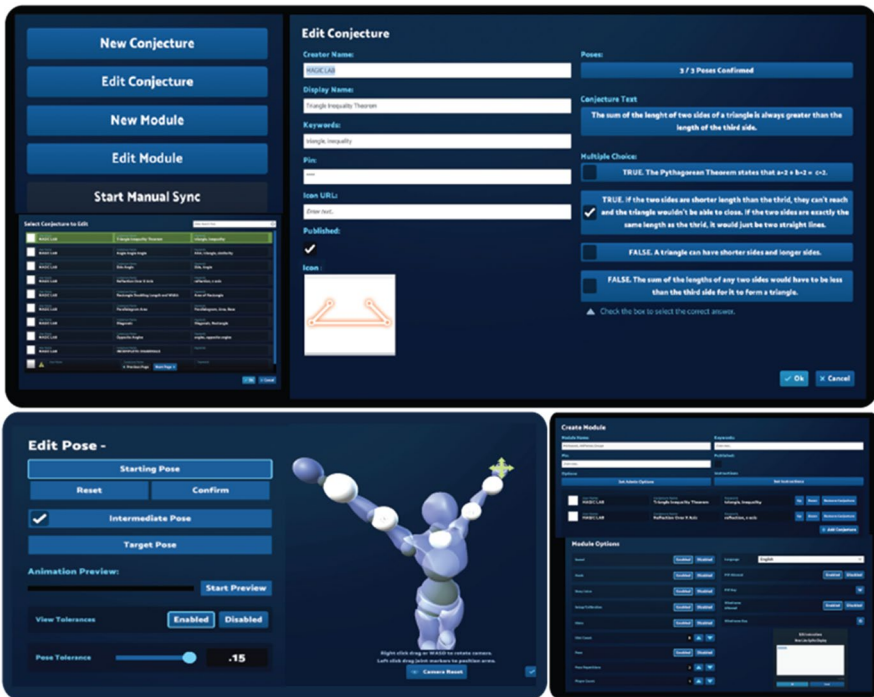


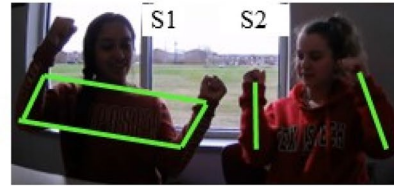
Fig. 3 The Conjecture and Module Editor: (top-left) Main Menu and Conjecture Selector; (top-right), Conjecture Editor; (bottom-left) The Pose Editor; (bottom-right) Module Builder and Admin Panel

1 S1: Okay. Um.

2 S2: Oh no.

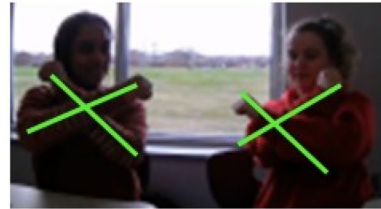
3 S1: Yeah, we could do something like that.

(arms bent at elbows and tilted to the side to represent a rhombus)



4 S1: And then that.

(arms crossed to represent the diagonals)



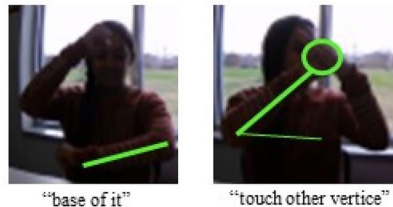
5 S2: But we need a third one though.

6 S2: Like this.

(left out straight, right arm slanted to form angle/triangle)



7 S1: Yeah. Try to like, 'cause this is the base of it. So it touches one of the vertices, and it would have to touch the other vertice.



“base of it”

“touch other vertice”

8 S2: Yeah.

9 S1: And it should be a reflection.

(pointing out the reflection on the understood rhombus)



“reflection”

10 S1: Like on the other side or whatever. Whatever that is.

(alternates between these two gestures, showing the two diagonals)

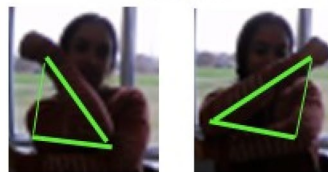


Fig. 4 Pair of high school students working together to come up with motions for the conjecture “The diagonals of a rhombus bisect the angles at all four vertices”

Students also used multimodal forms of communication and reasoning, including gesture, when they were playing through and solving problems in the THV, rather than creating their own poses. Figure 5 shows two students coming up with a justification as to why

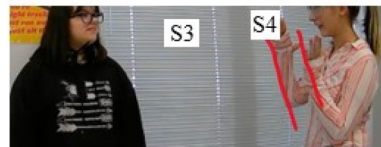
the conjecture “The area of a parallelogram is the same as the area of a rectangle with the same base and width” is true. We first see S4 using depictive gestures to display what a parallelogram looks like, with forearm gestures highlighting the two sets of parallel sides (Line 1). She uses vertical gestures that are slightly slanted, and horizontal gestures that are parallel to the floor. S3 then responds to these gestures by bringing into the exchange a new dynamic gesture, where the parallel vertical sides first go straight up and down, and then are transformed into a slanted position, as the horizontal parallel sides remain constant in their angle (Line 4). This transcript shows how learners build off of each other’s gestures and use gestures to dynamically transform simulated mathematical objects while engaging with embodied technologies. We again see that the most relevant mathematical knowledge in this excerpt is expressed via gesture rather than speech or writing alone. These analyses point to the importance of considering gesture as a means of knowledge sharing and knowledge demonstration that is on equal footing with other modalities.

MAET in immersive VR environments

We also engaged in a VR project where mathematics teachers, engaging in professional learning remotely at the beginning of the COVID-19 pandemic, used VR goggles in a collaborative simulation environment to learn about 3D geometric shapes (Huang et al., in press; Walkington et al., 2021). This VR environment allowed teacher-learners to join a virtual classroom as avatars (with their head, torso, and hands virtually rendered), and manipulate and interact with 3D shapes like cubes and cylinders. Teachers confronted scenarios where they had to explore conjectures about shapes or orient combinations of shapes or their own bodies in particular ways. We present MAET from this data that examines the ways in which the teachers coordinated multiple modalities in real time while in a VR environment, and the patterns by which they would activate modalities in sequence while learning. The teachers were in separate physical locations (i.e., their individual homes) wearing Oculus Quest VR goggles.

1. S4: A parallelogram is slanted and straight.

(arms bent at elbows with forearms tilted to the side to represent the diagonal sides of a parallelogram)
(arms bent at elbows with hands horizontally parallel to close the shape and represent the top and bottom of a parallelogram, moves parallel arms side to side three times to emphasize these are the top and bottom of the shape)



2. S3: Oh so it's like a. So. What it be the same? Or no?

3. S4: Wait isn't it talking about the area? I don't know.



4. S3: Yeah same as the area. Yeah I think it would because the tops are the same as the parallelogram and the sides just go like this. True.

(starts with arms vertically parallel to represent a rectangles sides, then slants them diagonally to the side to show the shift to a parallelogram)



Fig. 5 Two students formulating an argument as to why the conjecture “The area of a parallelogram is the same as the area of a rectangle with the same base and width” is true

Figure 6 shows two teachers, Amy and Lily, exploring a task about the volume of a cylinder, and how the volume relates to the radius and the height. An important element in this technology context is the existence of virtual objects that can be dynamically and collaboratively interacted with, which offers different affordances for gesture and action. Another important element to consider is that hands and bodies are rendered virtually, as is the surrounding world. In Line 1, Amy makes a gesture where she is showing what an increasing/decreasing action on the virtual object (the cylinder) would look like, without actually performing it. This gesture may function to connect her words to the objects in the environment and communicate her meaning to her partner. In Line 2, Lily responds with a beat gesture to emphasize her agreement. In Line 3 we see Lily move her virtual body forwards towards the cylinder, signaling that she is likely planning to interact with it, and then in Line 4 she carries out an action (resizing) on the virtual object. The limitations of the VR environment make it difficult to observe whether the learners' focus of attention is on their partner or on the object, given that they are positioned on opposite sides of the object. It can also be seen from the transcript that their hands are not always present—the hands have to be in the view of the VR goggles to appear.

In this episode, we observed an important multimodal sequence where students plan their actions with the virtual objects collaboratively first, using speech, gestures, and movements. The pair then begins to actually interact with the object afterwards to make observations. We named this theme “discussion before embodied action,” and examined the conditions of its presence across the corpus (see Huang et al., in press). Such sequences helped us to understand the interactions of different collaborative groups and the ways in which they approached different kinds of mathematical tasks. This can allow for better structuring of VR activities for collaboration.

MAET in shared augmented reality (shAR)

We also engaged in a shAR project where high school students explored conjectures about 2D and 3D geometric objects in pairs, using collaborative dynamic holograms projected by Microsoft HoloLens 2 goggles. The objects were rendered using the GeoGebra 3D software, and students were physically in the same location, with the objects rendered in an “anchored” manner such that they were in the exact same position for each learner. There were a variety of different simulations learners could interact with (see Fig. 7), and various different measurements and constructions could be dynamically added to each simulation with a blue panel that only the controller/facilitator could use. One affordance of the goggles is that we had access to first-person camera views for each learner, which we utilize in the present case.

We used MAET to share the ways in which students collaborated using these holograms and adjusted their actions to better align with the affordances of the technology (Walkington et al., in press). Figure 8 shows two students using an interactive hologram of parallel lines cut by a transversal (top photo Fig. 7) to explore the mathematical conjecture “If two parallel lines are cut by a third line, the pairs of corresponding angles are congruent.” The parallel lines are near eye-level for the students, and students are immersed in the parallel lines which are much larger than they are. This larger scale may allow for new types of mathematical exploration (Dimmel et al., 2021). For this reason, this transcript shows unique ways in which MAET can capture body movements and positionings related to mathematical objects. In Lines 1 and 2, we see the students adjusting the lines, with S6

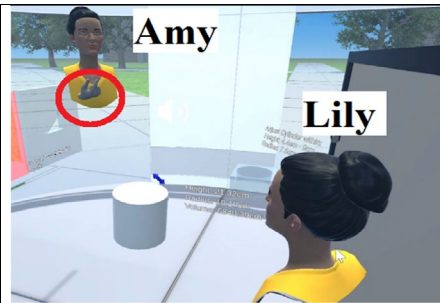
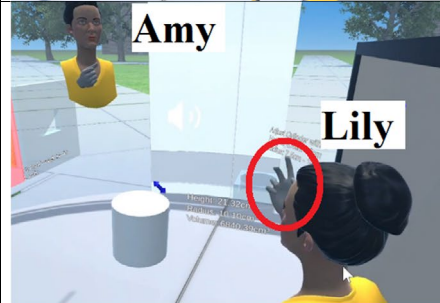
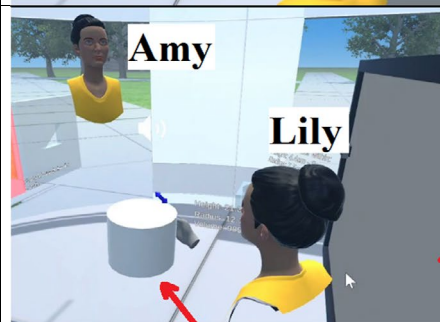

<p>1. Amy: Oh, so if the formula is radius squared so every time you increase it you increase it by the square.</p> <p><i>[Amy raises her right hand, open facing up, and then makes pinching motion with her index finger and thumb, and then moving it up and down]</i></p>	
<p>2. Lily: Yeah exactly.</p> <p><i>[Lily beats her right hand]</i></p>	
<p>3. Amy: Is that the formula? What was the formula for the volume?</p> <p><i>[Lily moves toward the cylinder]</i></p>	
<p>4. Lily: pi 'r' squared times...</p> <p><i>[Lily grabs side of the cylinder, and increases the radius of the cylinder]</i></p>	

Fig. 6 Amy and Lily in a VR environment exploring a task about whether increasing the radius or the height of a cylinder increases the volume more

trying to position his body between the lines. In Lines 3 and 4 we see the students engaging in Planning talk moves, with S6 maintaining his body's position between the lines as he moves closer to the transversal line. In Lines 5 and 6, we see the students tilting their heads and standing on their toes to get a new perspective on the set of parallel lines that they are immersed in. We coded body positioning across the corpus to see how learners

Simulation	Example Task/Conjecture	Simulation Photo
Parallel Lines	<p>If two parallel lines are cut by a third line, the pairs of corresponding angles are congruent.</p> <p>Is it true or false? Why?</p>	
Cylinders	<p>Given a cylinder with radius r and height h, the cylinder can be unrolled to include a rectangle with length h and width $2\pi r$.</p> <p>Is it true or false? Why?</p>	
Prisms	<p>If the length, width, and height of a cube are each doubled, then the volume increases by a factor of 8.</p> <p>Is it true or false? Why?</p>	
Pyramids or Cones	<p>The volume of a cone is one-third the volume of a cylinder with the same base and height.</p> <p>Is it true or false? Why?</p>	

Fig. 7 Example of GeoGebra 3D simulations, and accompanying conjectures, rendered using Microsoft HoloLens 2

move during different phases of reasoning (see Hunnicutt et al., under review). This MAET allowed for important observations about how scale, the positioning of the representation, and body positioning may be important design considerations in our context.

Significance and conclusion

These three cases together show MAET to be a flexible approach to understanding how learners engage with embodied, collaborative learning technologies. These kinds of analyses can lead to novel insights about designing technologies to leverage the distributed and embodied nature of cognition. They can also broaden what counts as “knowing” in academic settings that have traditionally privileged particular modalities and ways of exploring and communicating.

In this paper, we show how MAET can capture key collaborative moves, like collaborative gestures. These kinds of gestures, that arise through learners embodying ideas together in concert, have not been the focus of many previous analyses using this method. In addition, we add the element of interactions with dynamic, virtual objects in a technology-enhanced simulation environment, showing how scale, perspective, and collaborative actions on objects are important elements when considering multimodality in



Fig. 8 Two students exploring conjecture “If two parallel lines are cut by a third line, the pairs of corresponding angles are congruent”

these settings. This consideration of embodied collaboration practices and affordances of dynamic embodied technologies extends many previous illustrations of multimodal analysis.

Taken together, these three cases can help illustrate important factors that should be taken into account when using MAET. First, appropriately capturing detailed information about and analyzing how students collaborate over time across different modalities, as their reasoning develops, is key to this approach. Examining each interaction in the context

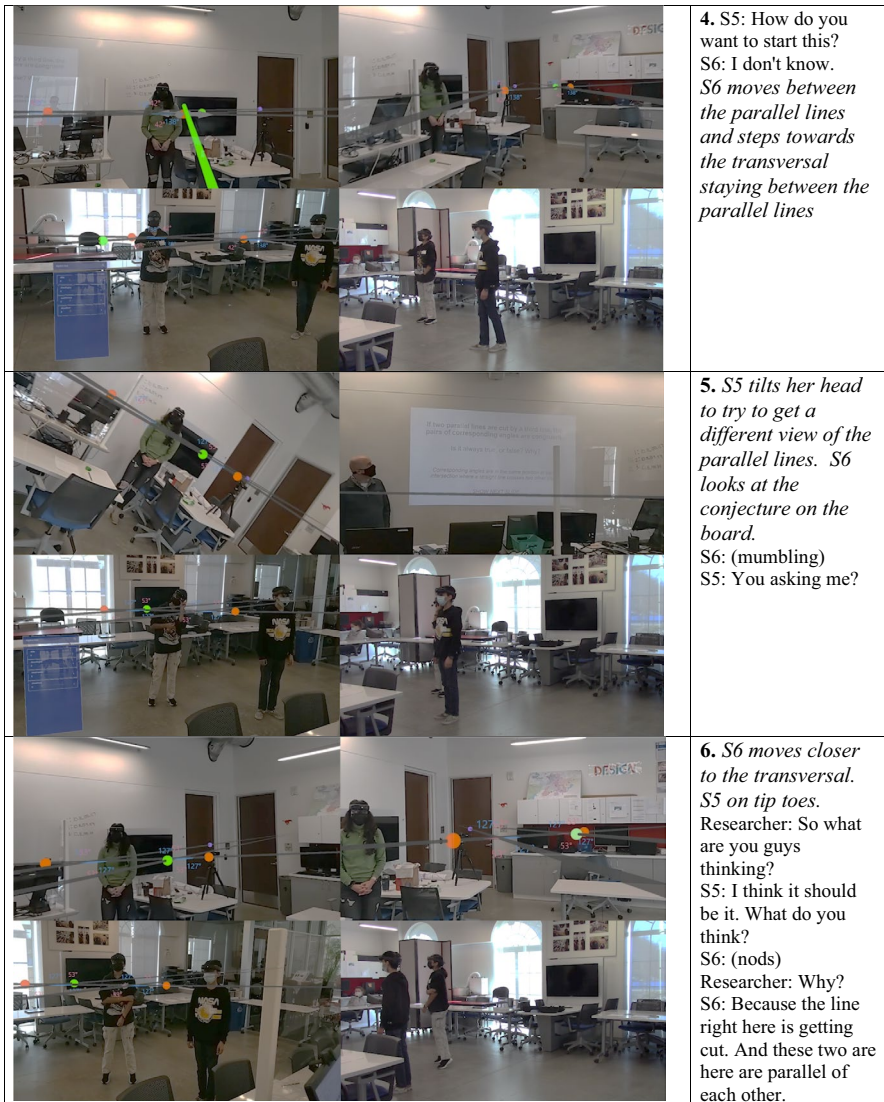


Fig. 8 (continued)

of historical interactions creates a detailed picture of how learners work together to build cohesion. Second, explicit consideration of gestures, actions and movements, and thinking through how these relate to targeted learning goals, is a key element of MAET which we illustrated. For example, students' developing understanding can be understood by the progression of their gestures, actions, and movements over time. This elevating of gestures and movements as forms of knowledge challenges traditional notions of "knowing" as being based on verbal and written responses. Third, different learning technologies will have different affordances for capturing multimodal data (e.g., body tracking, first-person recording, etc.), and the affordances of your system should be considered when planning

for different multi-modal streams. Fourth, the learning goals will influence the kinds of embodied and collaborative moves that unfold, and designing learning tasks for educational technologies with possibilities for embodied collaboration in mind may best leverage these technologies.

One weakness of the MAET method as presented here is its reliance on video camera feeds, rather than other forms of multimodal learning analytics (e.g., automatic gesture detection) that may be easier for researchers to collect in some contexts. In addition to such analytics being less intrusive and protecting learners' identities better than raw video/audio feeds, they may also be easier to quantitatively analyze—indeed, syncing and coding video data from multiple perspectives for micro-interactions can be quite time-consuming. We recommend that researchers consider which portions of their data may be richest and best suited for the application of MAET. Finally, we have found that a networked system of audio, motion, and video data capture technologies are necessary for MAET to be able to rigorously examine students' interactions across different modalities; the cost, technical expertise, and setup may be prohibitive in some cases.

By presenting three very different cases of using MAET with respect to technology platform (a motion capture game, a VR environment, and shared holographic AR simulations), we demonstrate the flexibility of this approach. We think this kind of analysis will be particularly beneficial as technologies begin to allow for collaboration and allow for learners to communicate with each other through ways other than typing, clicking, or talking. These issues are on the leading edge of current technology advancements, making MAET a highly effective method for studying these innovations.

MAET allows for an expanded view of what learners know and how they work together with technology, including their challenges and opportunities for providing important supports. Many researchers have found that when they start paying attention to gestures and other non-verbal forms of communication and reasoning, whole new ways of considering thinking and behavior are opened up. Researchers are able to learn more about human cognition and be better informed as they strive to develop effective designs of learning environments to promote understanding.

Acknowledgements The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A200401 to Southern Methodist University and grant R305A160020 to University of Wisconsin-Madison. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We would like to thank Michael Swart, Kelsey Schenck, Kasi Holcomb-Webb, and the members of our advisory board for their contributions to this research.

Funding Open access funding provided by SCEL, Statewide California Electronic Library Consortium.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abrahamson, D., Nathan, M. J., Williams-Pierce, C., Walkington, C., Ottmar, E. R., Soto, H., & Alibali, M. W. (2020). The future of embodied design for mathematics teaching and learning. *Frontiers in Education*, 5, 147.
- Abrahamson, D., & Sánchez-García, R. (2016). Learning is moving in new ways: The ecological dynamics of mathematics education. *Journal of the Learning Sciences*, 25(2), 203–239.
- Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language*, 44(2), 169–188.
- Alibali, M. W., & Nathan, M. J. (2012). Embodiment in mathematics teaching and learning: Evidence from learners' and teachers' gestures. *Journal of the Learning Sciences*, 21(2), 247–286.
- Andrews-Todd, J., & Forsyth, C. M. (2020). Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Computers in Human Behavior*, 104, 105759.
- Andrews-Todd, J., Jackson, G. T., & Kurzum, C. Collaborative problem solving assessment in an online mathematics task. *ETS Research Report Series*.
- Aukstakalnis, S., & Blatner, D. (1992). *Silicon mirage: The art and science of virtual reality*. Peach Pit Press.
- Bezemer, J., & Kress, G. (2015). Recognition. In J. Bezemer & G. Kress (Eds.), *Multimodality, learning and communication: A social semiotic frame* (pp. 1–14). Routledge.
- Blair, L. (2016). Understanding the differences between virtual reality, augmented reality and mixed reality. *Network World*, 1–4.
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 7–19.
- Deppermann, A. (2013). Multimodal interaction from a conversation analytic perspective. *Journal of Pragmatics: An Interdisciplinary Journal of Language Studies*, 46(1), 1–7.
- Dimmel, J., & Bock, C. (2019). Dynamic mathematical figures with immersive spatial displays: The case of handwaver. In G. Aldon & J. Trgalová (Eds.), *Technology in mathematics teaching mathematics education in the digital era* (pp. 99–122). Springer.
- Dimmel, J., Pandiscio, E., & Bock, C. (2021). The geometry of movement: Encounters with spatial inscriptions for making and exploring mathematical figures. *Digital Experiences in Mathematics Education*, 7(1), 122–148.
- Edwards, L. D. (2009). Gestures and conceptual integration in mathematical talk. *Educational Studies in Mathematics*, 70(2), 127–141.
- Enyedy, N. (2005). Inventing mapping: Creating cultural forms to solve collective problems. *Cognition and Instruction*, 23(4), 427–466.
- Gerofsky, S. (2007). “Because you can make things with it”: A rationale for a project to teach mathematics as a multimodal design tool in secondary education. *Journal of Teaching and Learning*, 5(1), 23–32.
- Göksun, T., Goldin-Meadow, S., Newcombe, N., & Shipley, T. (2013). Individual differences in mental rotation: What does gesture tell us? *Cognitive Processing*, 14(2), 153–162.
- Goldin-Meadow, S. (2005). *Hearing gesture: How our hands help us think*. Harvard University Press.
- Goldin-Meadow, S., & Beilock, S. L. (2010). Action's influence on thought: The case of gesture. *Perspectives on Psychological Science*, 5(6), 664–674.
- Goodwin, C. (2000). Gesture, aphasia, and interaction. *Language and Gesture*, 2, 84–98.
- Huang, W., Walkington, C., & Nathan, M.J. (in press). Coordinating Modalities of Mathematical Collaboration in Shared VR Environments. *International Journal of Computer-Supported Collaborative Learning*.
- Jewitt, C. E. (2017). *The Routledge handbook of multimodal analysis* (2nd ed.). Routledge/Taylor & Francis Group.
- Jewitt, C., & Henriksen, B. (2016). *Social semiotic multimodality*. De Gruyter.
- Johnson-Glenberg, M. C., Birchfield, D. A., Tolentino, L., & Koziupa, T. (2014). Collaborative embodied learning in mixed reality motion-capture environments: Two science studies. *Journal of Educational Psychology*, 106(1), 86–104.
- Kim, M., Roth, W. M., & Thom, J. (2011). Children's gestures and the embodied knowledge of geometry. *International Journal of Science and Mathematics Education*, 9(1), 207–238.
- Koschmann, T., & LeBaron, C. (2002). Learner articulation as interactional achievement: Studying the conversation of gesture. *Cognition and Instruction*, 20(2), 249–282.
- Lakoff, G., & Núñez, R. E. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being*. Basic Books.

- McNamara, D.S., Louwerse, M.M., Cai, Z., & Graesser, A. (2013). *Coh-Metrix version 3.0*. Retrieved from <http://cohmetrix.com>
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. The University of Chicago Press.
- Menache, A. (2011). Motion capture primer. In *Understanding Motion Capture for Computer Animation* (pp. 1–46). <https://doi.org/10.1016/b978-0-12-381496-8.00001-9>
- Milgram, P., & Kishino, F. (1994). A taxonomy of mixed reality visual displays. *IEICE Transactions on Information Systems*, *E77*(12), 1–15.
- Moll, H., & Tomasello, M. (2007). Cooperation and human cognition: The Vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society b: Biological Sciences*, *362*(1480), 639–648.
- Mondada, L. (2019). Contemporary issues in conversation analysis: Embodiment and materiality, multimodality and multisensoriality in social interaction. *Journal of Pragmatics*, *145*, 47–62.
- Nathan, M. J., & Alibali, M. W. (2021). An embodied theory of transfer of mathematical learning. In C. Hohensee & J. Lobato (Eds.), *Transfer of learning: Progressive perspectives for mathematics education and related fields* (pp. 27–58). Springer.
- Nathan, M. J., Schenck, K. E., Vinsonhaler, R., Michaelis, J. E., Swart, M. I., & Walkington, C. (2020). Embodied geometric reasoning: Dynamic gestures during intuition, insight, and proof. *Journal of Educational Psychology*, *113*(5), 929–948.
- Nathan, M. J., Walkington, C., Boncoddio, R., Pier, E. L., Williams, C. C., & Alibali, M. W. (2014). Actions speak louder with words: The roles of action and pedagogical language for grounding mathematical proof. *Learning and Instruction*, *33*, 182–193.
- Nathan, M. J., Wolfram, M., Srisurichan, R., Walkington, C., & Alibali, M. (2017). Threading mathematics through symbols, sketches, software, silicon and wood: Integrated STEM instruction to produce and maintain cohesion. *The Journal of Educational Research*, *110*(3), 272–293. <https://doi.org/10.1080/00220671.2017.1287046>
- Nathan, M.J., Walkington, C., & Swart, M. (2021). Investigating computer designs for grounded and embodied mathematical learning. In Rodrigo, M. M. T. et al. (Eds.) (2021). *Proceedings of the 29th international conference on computers in education* (pp. 237–246). Asia-Pacific Society for Computers in Education.
- Newcombe, N. S., & Shipley, T. F. (2012). Thinking about spatial thinking: New typology, new assessments. In J. S. Gero (Ed.), *Studying visual and spatial reasoning for design creativity*. Springer.
- Ng, O., & Sinclair, N. (2015a). Young children reasoning about symmetry in a dynamic geometry environment. *ZDM: International Journal on Mathematics Education*, *47*(3), 421–434.
- Ng, O., & Sinclair, N. (2015b). “Area without numbers”: Using touchscreen dynamic geometry to reason about shape. *Canadian Journal of Science, Mathematics and Technology Education*, *15*(1), 84–101.
- Norris, S. (2016). Concepts in multimodal discourse analysis with examples from video conferencing. *Yearbook of the Poznan Linguistic Meeting*, *2*(1), 141–165.
- Onyesolu, M. O., & Eze, F. U. (2011). Understanding virtual reality technology: Advances and applications. In M. Schmidt (Ed.), *Advances in computer science and engineering* (pp. 53–70). Intech Open.
- Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. LIWCNet.
- Pier, E. L., Walkington, C., Clinton, V., Boncoddio, R., Williams-Pierce, C., Alibali, M. W., & Nathan, M. J. (2019). Embodied truths: How dynamic gestures and speech contribute to mathematical proof practices. *Contemporary Educational Psychology*, *58*, 44–57.
- Roth, W. M. (2011). *Geometry as objective science in elementary school classrooms: Mathematics in the flesh*. Routledge.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction* (pp. 7–55). Academic Press.
- Salomon, G. (Ed.). (1993). *Distributed cognitions: Psychological and educational considerations*. Cambridge University Press.
- Shvarts, A., & Abrahamson, D. (2019). Dual-eye-tracking Vygotsky: A microgenetic account of a teaching/learning collaboration in an embodied-interaction technological tutorial for mathematics. *Learning, Culture and Social Interaction*, *22*, 100316.
- Sinclair, N. (2005). Chorus, colour, and contrariness in school mathematics. *THEN: Journal*, *1*(1), 1–15.
- Uttal, D. H., Miller, D. I., & Newcombe, N. S. (2013). Exploring and enhancing spatial thinking links to achievement in science, technology, engineering, and mathematics? *Current Directions in Psychological Science*, *22*(5), 367–373.
- Varela, F. J., Thompson, E., & Rosch, E. (2017). *The embodied mind* (revised). MIT press.

- Vest, N. A., Fyfe, E. R., Nathan, M. J., & Alibali, M. W. (2020). Learning from an avatar video instructor: The role of gesture mimicry. *Gesture*, *19*(1), 128–155.
- Vygotsky, L. S. (1978). *Mind in society*. Harvard University Press.
- Walkington, C., Chelule, G., Woods, D., & Nathan, M. J. (2019a). Collaborative gesture as a case of extended mathematical cognition. *Journal of Mathematical Behavior*. <https://doi.org/10.1016/j.jmathb.2018.12.002>
- Walkington, C., Gravell, J., & Huang, W. (2021). Using virtual reality during remote learning to change the way teachers think about geometry, collaboration, and technology. *Contemporary Issues in Technology and Teacher Education*, *21*(4), 713.
- Walkington, C., Nathan, M. J., Wang, M., & Schenck, K. (2022). The effect of cognitive relevance of directed actions on mathematical reasoning. *Cognitive Science*. <https://doi.org/10.1111/cogs.13180>
- Walkington, C. A., Nathan, M. J., Wolfram, M., Alibali, M. W., & Srisurichan, R. (2014). Bridges and barriers to constructing conceptual cohesion across modalities and temporalities: Challenges of STEM integration in the precollege engineering classroom. In J. Strobel, S. Purzer, & M. Cardella (Eds.), *Engineering in pre-college settings: Research into practice* (pp. 183–209). Purdue University Press.
- Walkington, C., Woods, D., Nathan, M. J., Chelule, G., & Wang, M. (2019b). Does restricting hand gestures impair mathematical reasoning? *Learning and Instruction*. <https://doi.org/10.1016/j.learninstruc.2019.101225>
- Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Candace Walkington is an Associate Professor in Mathematics Education and Learning Sciences at Southern Methodist University. She studies technology-enhanced interventions for secondary mathematics learning.

Mitchell J. Nathan is a Professor in the Department of Educational Psychology at University of Wisconsin-Madison. He studies cognitive processes related to STEM learning from a learning sciences perspective.

Wen Huang is a Postdoctoral Scholar at Southern Methodist University who studies VR technologies for STEM education.

Jonathan Hunnicutt is a Ph.D. student in the Simmons School of Education at Southern Methodist University who studies math teachers' adoption of technology.

Julianna Washington is a Ph.D. student in the Simmons School of Education at Southern Methodist University who studies technologies for math learning.