International Society of the Learning Sciences

Joint visual attention and collaboration in Minecraft

Chris Proctor, Dalia Antonia Caraballo Muller
chrisp@buffalo.edu, daliamul@buffalo.edu
University at Buffalo—SUNY

Abstract: In the service of a project organizing social futures (Penuel & O'Connor, 2018) in virtual worlds, this paper develops methods for analyzing collaboration in Minecraft through joint visual attention (JVA) and other multimodal learning analytics, and contributes a software package for doing so. We continuously calculate each player's position, gaze vector, and the block upon which their gaze falls based on server logs, revealing when players are looking at the same place at the same time. We show that JVA functions similarly in online worlds as offline, and provide examples of how JVA and other analytics can supplement qualitative analysis of online collaboration. In contrast to previous research on joint visual attention using dual eye-tracking glasses, JVA research in virtual worlds allows for larger collaborative groups, freedom of movement, and the opportunity to observe how players establish and negotiate genres of multimodal interaction through new forms of embodiment.

Introduction

More and more we live online. Even when disconnected from a phone or a computer, our offline lives increasingly refer to digital worlds to ground claims of authenticity and veracity rather than the other way around. Ubiquitous surveillance produces authoritative accounts of our behavior. QR codes linked to online databases attest that we have bought a ticket, hold a professional credential, or have been vaccinated. And online discourse plays a central role in in determining the visibility and social meanings of offline events. Baudrillard's (1981) image of people living in a map of a world which no longer exists is a poignant description of the generation of Zoomers coming of age during the COVID-19 pandemic. There is no "real" against which their syncretic digital lives might correspond and be measured.

The synthetic structure of our realities holds potential for freedom as well as oppression, depending on which futures we orient toward and who has the power to realize them. The present study is part of a broader project to develop pedagogies of organizing social futures (Penuel & O'Connor, 2018) through what Glissant (1990) calls *Relación*, or a focus on the whole rather than on the individual, seeing ourselves in and through relation with others. For Glissant, *Relación* is born out of the pain and dislocation of the slave trade. An unexpected benefit of this tragedy is the possibility of building identities and cultures disentangled from territorial nationalism and colonial violence. Without generalizing the historical specificity of the Middle Passage or the experiences of those who live in its wake, our goal is to develop inclusive learning environments in which the collaborative production of realities opens the way for critical imagining.

Drawing on CSLC research on computer-mediated intersubjectivity and collaboration, the present study is a first step toward these goals. In the context of a summer Minecraft workshop for fifth-grade students (average age of 10), we demonstrate the feasibility of collecting joint visual attention data in open-ended collaborations and analyzing them together with other multimodal learning analytics (Blikstein, et al., 2016). We present two contrasting cases of groups engaged in successful and unsuccessful collaboration. We close with a discussion of how these tools will support basic research on multimodal sense-making, collaboration, as well as how they will support the iterative design of our pedagogical approach.

Background

Intersubjectivity and collaboration in CSCL

CSCL has long been interested in how computer-mediated intersubjectivity might create the conditions for more effective collaboration. While CSCL encompasses many theoretical orientations, Akkerman and colleagues' (2007) review identified two dominant paradigms, one cognitive/individualistic and the other socio-cultural. The paradigms can be distinguished by the extent to which they consider the individual to be separable from social context. Both paradigms frame collaboration as "a process of building and maintaining a shared conception of a problem" (Akkerman, et al., 2007, p. 39). Intersubjectivity (sometimes called common ground, shared understanding, or collective mind) is regarded as an essential basis for the shared conception of a problem. In designing for *Relación,* our long-term goal is an enriched understanding of collaboration which also

includes a shared conception of each others' epistemological and ontological status. The "problem" on which the group's collaboration is focused is the group's own intersubjectivity and how they are going to work together. Our project is aligned with recent work in computing education extending cognitive and situated framings of computational thinking to include a broader, critical framing (Kafai, Proctor, & Lui, 2019).

Research on collaboration relies on measures of collaboration quality and analytics capturing aspects of process. Meier, Spada, and Rummel's (2007) widely-adopted rating scheme, used in this study, defines collaboration in terms of nine dimensions of process: sustaining mutual understanding, dialogue management, information pooling, reaching consensus, task division, time management, technical coordination, reciprocal interaction, and individual task orientation. Analytics of collaborative learning have also been used as quantitative measures of collaboration quality, though Wise, Knight, and Shum (2021) note tension between this approach (often unsupervised and atheoretical) and the use of analytics to surface patterns to be used in qualitative analysis of collaborative processes. With this tension in mind, the new forms of analytics presented in this paper are presented both as aides to qualitative analysis and as potential quantitative measures.

One modality of analytics, joint visual attention (JVA), has received particular attention the last decade as a mechanism by which groups establish intersubjectivity. Laboratory studies with dual eye-tracking cameras have demonstrated the importance of JVA in coordinating shared meaning-making. JVA has been correlated with more accurate communication (Richardson & Dale, 2005), more effective collaboration (Schneider et al, 2019), and with collaborative learning gains (Olsen et al., 2020). Interventions designed to support more effective JVA have been shown to improve performance on collaborative tasks (Schneider, et al., 2016). More recent research has moved away from using JVA as a singular measure of collaboration, instead treating JVA as one of multiple modalities which are recruited together in establishing intersubjectivity (Schneider, et al., 2018; Olsen, et al., 2020; Sharma, et al, 2021).

Collaboration in Minecraft

Minecraft is a multiplayer online game in which players inhabit and modify a blocky, pixellated world (see Figure 1). Supported by third-party extensions to the Minecraft server and client software (*mods*), community-hosted servers offer a variety of social forms such as games, contests, storytelling, education, and activism. Although it is evident that players of Minecraft engage in extensive in-game collaboration (Mørch & Eielsen, 2021), it is unclear whether JVA plays the same coordinating role in Minecraft or other immersive online worlds that it plays in offline collaboration. While playing Minecraft, players are embodied first in their corporeal bodies and secondly in their in-game avatars. Visual attention has two layers: first, how a player's gaze moves across the screen, and second how the player orients her avatar to look around the world. Prior JVA research has tracked eye saccades, or quick darting shifts of focus around the visual field. In this research we assume the player's ocular gaze is fixated on the center of her screen, and that she relies entirely on moving her avatar to look around. This is not an outrageous assumption: one research assistant who is an avid gamer affirmed that, in contrast to first-person shooter games where situational awareness is paramount, it is common in Minecraft to remain fixated on the center of the screen, where a small crosshairs orients the eye to the location where block-building or block-removing actions will have their effect.

If JVA does function similarly in Minecraft as in other contexts, a potential advantage to online collaboration research is greater ecological validity (in an admittedly synthetic ecology). No further instrumentation is needed beyond the standard Minecraft client and server to precisely and continuously capture players' visual attention. Players are free to move around, manipulate their environment, and collaborate in groups larger than the dyads typically engaged in JVA research. The artifice of Minecraft could actually be helpful for research, as modalities of communication (e.g. gaze, gesture, movement and position, discourse, the manipulation of artifacts) can also be perfectly captured by the research setup. Given the limited articulation of avatars compared to the human body, face, and eyes, it may be easier to study the emergence of genres of multimodal interaction as players figure out how to work with what they have.

Research Questions

1. **Do high-collaboration groups in Minecraft engage in more joint visual attention?** We hypothesize that groups which collaborate more successfully will also exhibit more joint visual attention.
2. **Does joint visual attention interact with other modalities in Minecraft? Is coordination between modalities dependent on effective collaboration?** We hypothesize that the rate of block activity (placing and removing blocks) will be different in JVA and non-JVA intervals, with a greater difference in high-collaboration groups.

Methods

Context

This study was conducted within an eight-week summer workshop for ten-year-old children focused on collaboration in Minecraft. The ten participants were recruited through existing social ties; all were experienced players of Minecraft. Two undergraduate researchers participated in the workshops and collected fieldnotes alongside the paper's authors. The workshop met for 75 minutes once per week. Another hour of supervised play was available each week, and participants were additionally free to log in at other times.

The workshop was conducted entirely online during summer 2021. Each participant would join the workshop's Minecraft world and also log in to a Zoom meeting which allowed for voice communication. The previous school year had been almost entirely online due to the COVID-19 pandemic, so participants and their families had experience and established practices for attending classes and socializing online. Participants logged in using a variety of devices, including desktop computers, laptop computers, iPads, and Nintendo Switch.

In the early weeks of the workshop, participants were placed in groups of two or three and asked to engage in open-ended collaboration, deciding together what to do and how to do it. The collaborative episodes studied in this paper are drawn from weeks 1, 2, and 4, in which groups were given the following challenges:

- Week 1: Build a single structure which represents everyone in your group.
- Week 2: Build a prototype structure for the world's *spawn point* (the starting location) which represents everyone in the workshop and the qualities we want for this world.
- Week 4: Build a prototype of a structure or object which creates a sense of unity and inclusiveness for our world.

Figure 1. Screen shots from group 1 (left; high-collaboration) and group 3 (right; low-collaboration)

Figure 1 shows screenshots from groups 1 and 3 in Week 1, contrasting cases of successful and unsuccessful collaboration (as defined in Meier, Spada, and Rummel's (2007) collaboration measure) to which we shall return in each of the subsequent analyses. These cases will be used to illustrate the utility of the analytical tools developed in this paper. In a subsequent publication we plan to fully analyze these cases using the tools developed here.

Group 1, shown on the left side of Figure 1, decided early in their collaboration to build a three-story house with one room for each group member, and spent much of the rest of the session considering ideas for the structure, such as how tall each story should be and what form the roof should take. In Figure 1 (left), the group is working together to build the outer walls of their house, frequently alternating between building and observing other group members' work. One member, BobOriginal, plays a key role in managing the group's dialogue by affirming group members' contributions and connecting ideas to the emerging consensus.

Group 3, shown on the right side of Figure 1, had less successful collaboration. Dialogue in this group was much more complex, with one high-status member, Mutton, enthusiastically engaged in the task while dominating the dialogue. Despite their enthusiasm, they seldom acknowledged ideas presented by others and frequently talked over them. Over the course of the session, the other two group members become increasingly

frustrated. The right side of Figure 1 shows a moment late in the segment, when collaboration has broken down and each group member is building their own structure.

Data sources

Minecraft presents players with a shared virtual world synchronized by a central server with updates streamed to each client. The server was running a customized version of the SuperLog mod (Andross, 2020) which logged every in-game event, including placing and removing blocks, player movement, and changes in player gaze. The vector of a player's gaze was projected out from the avatar's in-game eye location until it hit a solid block, allowing us to additionally log the coordinates of the player's target block, or what the player saw at the center of her screen. Over the eight weeks of the workshop, we logged over six million events.

The researchers' Minecraft clients were running the Minecraft Replay mod (CrushedPixel & johni0702, 2021), which captures a continuous stream of state data arriving from the server. Having captured a replay, researchers could later walk around in the world, pan across time, and render camera paths through the captured time and space as videos. Additionally, we collected audio recordings and transcripts from Zoom and additional audio recordings from software on researchers' computers. (Small groups often worked together in Zoom breakout rooms, which Zoom does not record.)

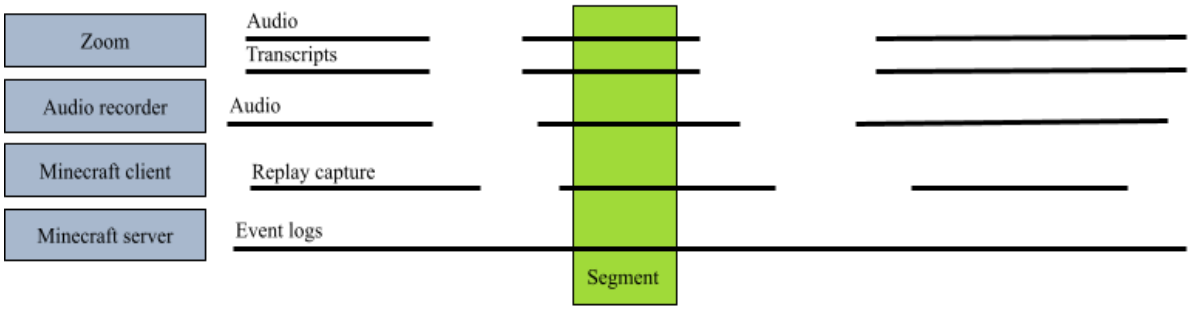

Figure 2. Schematic diagram showing a segment slicing across multiple media streams.

Schneider, et al. (2018) noted that synchronizing multiple media streams for analysis presents a major logistical challenge. One of this paper's methodological contributions is a software framework for multimodal learning analytics research, released as a git repository at (BLINDED URL). First, the software collects metadata for each media stream via data entered by researchers as well as from the file itself when possible. This metadata includes the start time for the stream in universal coordinated time (UTC). Then, the researcher can specify a segment of interest bounded by UTC start and stop times, as well as one or more products to be produced for that segment. The software calculates offsets for each media stream and processes them according to the implementation of the products requested. Examples of segment products include videos made by combining audio and video streams (the basis for subsequent qualitative analysis), datasets produced from log files, and the diagrams shown in Figures 3, 4, 5, and 6. The result is a flexible and reliable system in which views into the data can be quickly produced and adjusted via declarative configuration files.

Analysis

To answer the first research question, we began by identifying episodes of collaboration in the first half of the workshop, when participants were assigned to small groups and asked to complete a task together. We cropped these segments so that they started and ended with active collaboration, excluding from analysis time it took for groups to get started. After identifying the temporal bounds of each segment, we produced videos documenting each collaboration session. We reviewed these videos as well as workshop fieldnotes and research team discussions to classify each group as "high collaboration" or "low collaboration" using Meier, Spada, and Rummel's (2007) collaboration measure.

For a measure of JVA, we adapted Schneider and Pea's (2013) approach in which each moment of collaboration is considered to have joint visual attention if the euclidian distance between the two group members' gaze has been less than some threshold within the previous or subsequent two seconds. (The logs sometimes recorded multiple events per second; we used a granularity of one second and defined each second as having JVA if it contained any moment of JVA.) We used a distance threshold of 6 blocks. Like Schneider et al. (2016), we found that results were not very sensitive to changes in time and distance threshold parameters.

In order to extend previous JVA measures to group sizes larger than two, we decided to consider the group member the unit of analysis, and to consider a group member to be engaged in JVA whenever their gaze coincided with any other group member. The ten workshop participants were grouped differently each week (8 different groups across all three weeks), resulting in a total of n=19 group-memberships. The means and standard deviations of percentage of time group members spent in JVA was comparable for members of dyads and for groups of three. For the JVA visualization shown in Figure 3, we decided to present pairwise JVA for each combination of two members in the group, as knowing which members are sharing JVA at a particular moment has interpretive value. With measures for JVA and collaboration in hand, we applied a one-tailed Welch's unequal variances t-test (α=0.05) to test whether the percentage of groupwork time spent in JVA was higher in high-collaboration groups than in low-collaboration groups.

To answer the second research question, we extended our analysis of JVA and collaboration to include the modality of block activity, or placing and removing blocks from the environment. We extracted block activity events from the server logs and aligned them with JVA for each segment. Following Jermann, et al. (2011) and Schneider, et al. (2018), we used cross-recurrence diagrams to illustrate how a dyad's visual attention at different points in the collaboration coincided spatially. We augmented these diagrams with histograms showing block activity.

In order to test the hypothesis for the second research question, we used the same JVA measure as in RQ1, but separated periods in which the group member engaged in JVA with any other group member from periods without JVA. Using the logs discussed in the previous paragraph, we divide the number of block actions in each period by its duration so that block activity is expressed in terms of block actions per second. For each group member, we then calculate the difference in average block activity during JVA and during non-JVA periods. We applied a two-tailed Welch's unequal variances t-test (α=0.05) to test whether the difference in block activity between JVA and non-JVA periods was different for members of high-collaboration groups and for low-collaboration groups.

Results

The first research question asked whether high-collaboration groups engage in more JVA. Figure 3 plots pairwise joint visual attention. These plots clearly align with the descriptions of each group's collaboration given above, which were based on video reconstructions of collaboration and fieldnotes. In group 1 (high-collaboration), there are several early moments where all three group members share visual attention. BobOriginal appears to play a leading role, engaging in JVA with both partners for roughly the first third of the collaboration, then alternating JVA with Zora and FallWhistle. This pattern corresponds to BobOriginal's role managing dialogue, pooling resources, and helping the group reach consensus, discussed above. In contrast, there was no time in group 3's collaboration in which all three members shared visual attention. Mutton's pattern of JVA is comparable to BobOriginal's in group 1, engaging sequentially with fellow group members, who have very little JVA with each other. However, whereas BobOriginal's JVA aligned with managing group 1's dialogue, Mutton's JVA corresponds to their dominance of group dialogue. These plots will be important artifacts guiding later qualitative analysis of how visual attention and dialogue interact in collaboration.

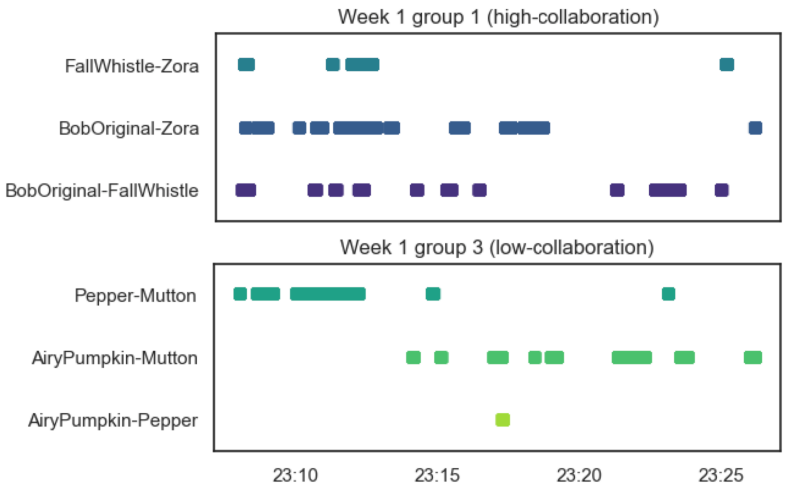

Figure 3 Joint visual attention plots for groups 1 and 3.

Figure 4 shows the average percentage of time spent in JVA for members of high-collaboration groups was nearly twice that of low-collaboration groups, confirming our hypothesis. The t-test results, $t=2.03$; $p < 0.031$, confirm that the difference is statistically significant.

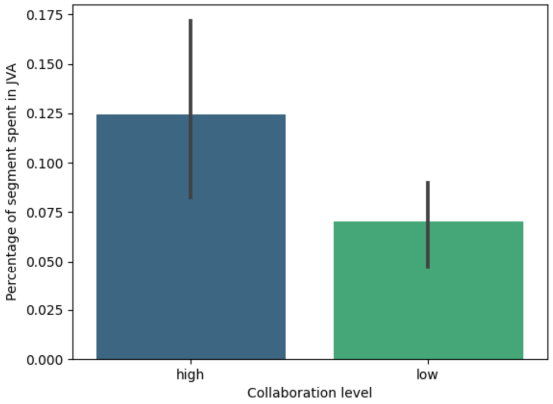

Figure 4. Percentage of time spent in JVA for members of high- and low-collaboration groups.

The second research question asked how JVA interacts with other modalities in Minecraft collaboration. Cross-recurrence plots could support a qualitative answer to this question, showing asynchronous spatial alignment of group members' gaze. That is, a cross-recurrence plot shows JVA as well as when one group member is looking at a point where the partner looked in the past or where they will look in the future. The large square in Figure 5 below presents a cross-recurrence plot for BobOriginal and Zora from group 1. A black pixel in the image at position (x, y) means that where BobOriginal was looking at time x is the same location (within a distance threshold) as where Zora was looking at time y. Black pixels along the diagonal line from the bottom left to the top right represent moments of JVA. Black pixels below the diagonal represent when BobOriginal looked somewhere Zora had looked previously, and vice versa.

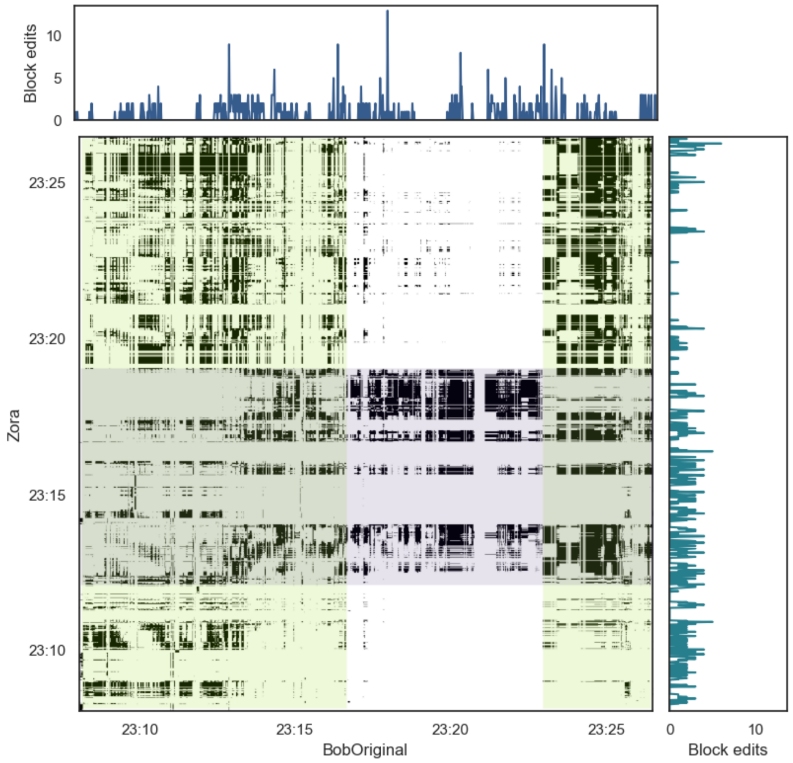

Figure 5. Cross-recurrence plot augmented with a histogram of each player's block activity.

The green-highlighted vertical bars in Figure 5 represent intervals in which BobOriginal looked at locations where Zora was looking for most of the segment. Similarly, the purple-highlighted horizontal bar

represents Zora looking at locations where BobOriginal was looking for the middle portion of the segment. These patterns suggest that these two players (within their group of three) each spent most of the segment looking at one location, with occasional visits to the other.

We augmented the cross-recurrence plot with histograms on the top and right showing BobOriginal's and Zora's block edits per second. Plotting JVA and block activity together makes it possible to observe relational patterns between joint attention and block activity. Zora had a steady rate of block activity until 23:19, which coincides with Zora's visual attention shifting to the location where BobOriginal was looking at the beginning and the end of the segment. Indeed, our video reconstruction of the session shows that at this point Zora shifted from working on building the floor of a room to discussing the structure of the walls with BobOriginal. BobOriginal's block activity appears to spike at transition points which are also visible on the cross-recurrence plot (23:13; 23:17, and 23:22). We will need to integrate dialogue into JVA and block activity in order to understand whether BobOriginal's increased activity was stimulated by shifts in Zora's visual attention, whether Zora's attention was drawn by BobOriginal's activity, or whether there is another explanation. We plan to investigate these patterns further in future research.

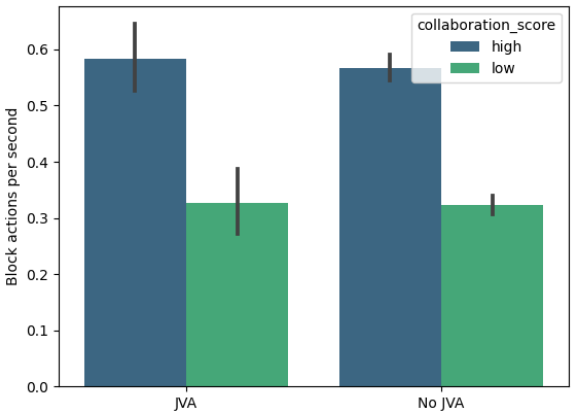

Figure 6. Average block actions per second for members of high-collaboration and low-collaboration groups, during JVA and non-JVA intervals.

Our hypothesis for RQ2 was that joint visual attention plays a role in coordinating action for high-collaboration groups but not for low-collaboration groups. If this were the case, the difference in block activity between JVA and non-JVA intervals would be greater in high-collaboration groups than in low-collaboration groups. However, Figure 6 shows that this is not the case. Although high-collaboration groups and low-collaboration groups have very different levels of block activity, JVA does not appear to have an effect on block activity levels in either case. Considering the patterns observed in analyzing Figure 5, we suspect that there may be more subtle interactions between JVA and block activity not captured by this analysis.

Discussion

This paper demonstrates the feasibility of analyzing JVA alongside other multimodal analytics in Minecraft, and contributes software tools which may be of use to other researchers. The results from RQ1 suggest that JVA plays an important role in collaboration in Minecraft just as it does offline. While the results for RQ2 do not support our hypothesis that effective collaboration moderates a relationship between JVA and block activity, we inadvertently observed a sizable gap in block activity between high-collaboration and low-collaboration groups. Our brief qualitative analysis of Figure 5 suggests that there may be interactions between JVA and block activity not modeled in our hypothesis for RQ2.

Beyond the novel context of Minecraft, this paper extends research on JVA and collaboration to groups larger than dyads and to three dimensional space in which players are free to move around. Neither of these have been previously reported, and they hold promise as more robust contexts for analysis of the multimodal mechanisms of collaboration. For example, recent JVA research has examined leadership roles in guiding a group's visual attention (Schneider, et al., 2018). Extending the group size from two to three could increase the visibility of participant roles by creating contexts where one member may respond to interaction between other group members. Extending JVA research into contexts where participants can move around, inhabit, and reconfigure the world around them could also support the integration of research on multimodal collaboration with research on the social construction of place. Finally, exploring how subjectivity and intersubjectivity

emerge through multimodal sensemaking will support our research on *Relación*, or the work of building and maintaining a shared conception of intersubjectivity from which social futures can be organized.

Conclusion

This paper develops methods which support our larger project of organizing educational experiences which meet the needs and potentials of the generations coming of age during and after the COVID-19 pandemic, for whom computing is the infrastructure of private embodied realities as well as social worlds. In future work, we intend to use the methods presented here to more deeply analyze multimodal collaboration in virtual worlds and to use them in our design-based research developing critical pedagogies.

References

Akkerman, S., Van den Bossche, P., Admiraal, W., Gijselaers, W., Segers, M., Simons, R.-J., & Kirschner, P. (2007). Reconsidering group cognition: From conceptual confusion to a boundary area between cognitive and socio-cultural perspectives? *Educational Research Review*, 2(1), 39–63.

Andross (2020). *SuperLog*. (Version 1.2) [Computer software]. https://superlog.andross.fr/

Baudrillard, J. (1995). *Simulacra and simulations* (S. Glaser, Trans.). University of Michigan Press.

Blikstein, P., & Worsley, M. (2016). Multimodal Learning Analytics and Education Data Mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, 3(2), 220–238.

CrushedPixel & johni0702 (2021). *Minecraft replay*. [Computer software]. https://www.replaymod.com/

Glissant, É. (1997). *Poetics of relation* (B. Wing, Trans.). University of Michigan Press.

Jermann, P., Mullins, D., Nüssli, M.-A., Dillenbourg, P., & Nuessli, M.-A. (2011). Collaborative Gaze Footprints: Correlates of Interaction Quality. *Proceedings of the 9th International Conference on Computer Supported Collaborative Learning*, 1, 184–191.

Kafai, Y. B., Proctor, C., & Lui, D. (2019). From Theory Bias to Theory Dialogue: Embracing Cognitive, Situated, and Critical Framings of Computational Thinking in K-12 CS Education. *Proceedings of the 2019 ACM Conference on International Computing Education Research*, 101–109.

Meier, A., Spada, H., & Rummel, N. (2007). A rating scheme for assessing the quality of computer-supported collaboration processes. *Int'l Journal of Computer-Supported Collaborative Learning*, 2(1), 63–86.

Mørch, A. I., Eielsen, C. S., & Mifsud, L. (2021). Using Minecraft to Reconstruct and Roleplay Local History: Intersubjectivity, Temporality, and Tension. In C. Hmelo-Silver, B. De Wever, & J. Oshima (Eds.), *Reflecting the Past and Embracing the Future: Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning* (pp. 27–34). ISLS.

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

Penuel, W. R., & O'Connor, K. (2018). From Designing to Organizing New Social Futures: Multiliteracies Pedagogies for Today. *Theory Into Practice*, 57(1), 64–71.

Richardson, D. C., & Dale, R. (2005). Looking To Understand: The Coupling Between Speakers' and Listeners' Eye Movements and Its Relationship to Discourse Comprehension. *Cognitive Science*, 29(6), 1045–1060.

Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., & Pea, R. (2016). Using Mobile Eye-Trackers to Unpack the Perceptual Benefits of a Tangible User Interface for Collaborative Learning. *ACM Transactions on Computer-Human Interaction*, 23(6), 1–23. https://doi.org/10.1145/3012009

Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., & Pea, R. (2018). Leveraging mobile eye-trackers to capture joint visual attention in co-located collaborative learning groups. *International Journal of Computer-Supported Collaborative Learning*, 13(3), 241–261.

Sharma, K., Olsen, J. K., Verma, H., Caballero, D., & Jermann, P. (2021). Challenging Joint Visual Attention as a Proxy for Collaborative Performance. *Reflecting the Past and Embracing the Future: Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning*, 91–98.

Wise, A. F., Knight, S., & Shum, S. B. (2021). Collaborative learning analytics. In U. Cress, C. Rosé, A. F. Wise, & J. Oshima (Eds.) *International handbook of computer-supported collaborative learning*. (pp. 371-388). Springer.

Acknowledgements

We are grateful to the parents and children who participated in this research. Isabelle Ondracek and Robert McManus served as undergraduate research assistants.