International Journal of Research & Method in Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/cwse20

The modality of mathematics lesson observations: comparing the results of live and video-based coding

F. Paul Wonsavage, Samuel Otten, Amber G. Candela & Zandra de Araujo

To cite this article: F. Paul Wonsavage, Samuel Otten, Amber G. Candela & Zandra de Araujo (22 May 2024): The modality of mathematics lesson observations: comparing the results of live and video-based coding, International Journal of Research & Method in Education, DOI: 10.1080/1743727X.2024.2350068

To link to this article: https://doi.org/10.1080/1743727X.2024.2350068

Published online: 22 May 2024.

INTERNATIONAL JOURNAL OF RESEARCH & METHOD IN EDUCATION
https://doi.org/10.1080/1743727X.2024.2350068

Routledge
Taylor & Francis Group

The modality of mathematics lesson observations: comparing the results of live and video-based coding

F. Paul Wonsavage [a], Samuel Otten [b], Amber G. Candela [c] and Zandra de Araujo [d]

[a]Lastinger Center for Learning, University of Florida, Gainesville, FL, USA; [b]Learning, Teaching, and Curriculum, University of Missouri, Columbia, MO, USA; [c]Department of Educator Preparation and Leadership, University of Missouri – St. Louis, St. Louis, MO, USA; [d]Lastinger Center for Learning, University of Florida, Gainesville, FL, USA

ABSTRACT

Classroom observations are an integral part of qualitative educational research. Traditionally, classroom observations have been done in-person, with one or more researchers being physically present in a classroom to observe and take field notes. With the proliferation of video technology, researchers are now able to conduct classroom observations at a distance using a variety of technological tools. When deciding on whether to conduct live or video observations, researchers must consider how the observation modality may influence the data. We address this consideration by independently conducting both live and video observations of the same mathematics classroom lessons to identify similarities and differences in the resultant coding between the two modalities. Our findings suggest there are some dimensions of mathematics classroom instruction unaffected by the live or video modality (e.g. nature of discourse, student contribution length) and others that are affected (e.g. lesson connections, mathematical development). Thus, when making decisions about observation modality, it is important to consider the focus of one's inquiry in addition to other factors such as cost, human capacity, geographical location, and more.

ARTICLE HISTORY
Received 8 December 2023
Accepted 7 April 2024

KEYWORDS
Classroom observations; mathematics teaching; video analysis; qualitative methodology

Introduction

Classroom observations can make teacher and student interactions visible (e.g. Star and Strickland 2008, Ayuwanti et al. 2021), unlike other data sources (e.g. student achievement data, interviews, surveys), and as such, classroom observations privilege indirect measures of instructional quality (Bostic et al. 2021). Researchers often use classroom observations not only to study teachers' instruction and interactions with and among students (e.g. Ing and Webb 2012, Boston et al. 2015), but also for purposes of professional development, evaluation, data triangulation, and more.

Classroom observations often occur by taking in-person field notes or video recording the lesson. With the proliferation of technology and in the aftermath of the COVID-19 pandemic, the decision to use video observations in place of live observations has become increasingly popular (Erickson 2006, Mac Mahon et al. 2019, Gold and Windscheid 2020, Dockerty 2022, Ramakrishnan 2023), sometimes without the researcher even being present to operate the camera. The prevalence of video observations has led researchers to ask questions such as, how live observation compares to video (Gridley et al. 2018), do the two modalities yield similar inferences about teaching (Casabianca et al. 2013), and whether are there important differences between the two modalities (Curby et al. 2016)? Methodological approaches aimed at answering these questions have been largely

CONTACT Paul Wonsavage wonsavagef@coe.ufl.edu 0711 Norman Hall, PO Box 117052, Gainesville, FL 32611-7052, USA

quantitative, focused on reliability and validity between the two modalities. Gridley and colleagues (2018), although not focused on classroom lessons, addressed similar matters by observing parent–child interactions within the home and did not find significant differences between coding live or through video. In contrast, Casabianca and colleagues (2013) and Curby and colleagues (2016) both found, in observations of algebra and preschool classrooms, respectively, that live coding led to slightly higher scores than video. Although these studies' contexts and results varied, all three noted live and video modalities are not interchangeable. Moreover, they provide evidence to suggest, that when selecting a data collection modality, the decision to observe live or through video should not be taken lightly.

The decision can be informed by researchers who have explored the advantages and disadvantages of video and live observations. For instance, conducting live observations may provide a fuller sense of the classroom context (Casabianca et al. 2013, Curby et al. 2016), but involves potentially burdensome travel, imprecise time tracking of classroom events, and missing subtle complexities within classrooms (Grossman 2014). Video observations allow viewers to stop and replay moments of particular interest or notice things on camera they may have missed upon first viewing, but video observations rely on a recording device and its successful operation (e.g. battery, field of vision, audio) (Ryan 1995, Lemke 2007, Haidet 2009, Blikstad-Balas 2017, Ing and Samkian 2018). Technological innovations such as increased storage capacities, video devices that track the teacher (e.g. Swivl), and omnidirectional cameras that capture a 360-degree viewing angle have facilitated remote observations of classrooms (Chilton and McCracken 2017, Mac Mahon et al. 2019, Ferdig and Kosko 2020). We anticipate the ease and quality of classroom video observation further improving with time, but it remains important to consider if and when one modality, video or live, might be better suited than the other. It is not a question of which modality is better, but rather, which one will more fully capture the phenomena under study given one's capacity, resources, and other constraints? The purpose of our study was to explore the methodological considerations involved with live and video classroom observations more deeply in the context of mathematics teaching.

Background literature

Classroom observations, whether live or video-recorded, are a critical source of data when researchers study classroom instruction (e.g. Borko et al. 2005). Given the complexity of instruction, researchers who seek to study it must make a number of decisions regarding which aspects of instruction are of greatest import to their study design. Researchers might seek to answer research questions related to mathematics classroom instruction broadly, attending to the time spent on various instructional activities. For instance, in-person observations have been used to identify and compare the percentages of class time spent on various instructional activities and formats (Weiss 2003, Grouws 2010). Others who studied mathematics classroom instruction broadly have used observation tools, such as the Reformed Teaching Observation Protocol ([RTOP], Sawada 2002), to get an overall sense of the extent to which mathematics instruction at a school or district level is reform-oriented (e.g. Adamson et al. 2003; Amrein-Beardsley et al. 2012). The RTOP is suitable for live and video observations (Boston et al. 2015).

Studying mathematics classroom instruction can also be done by focusing on the mathematical content of one's lesson. Researchers have dug deeply within this dimension using both live and video observations. An example is the Instructional Quality Assessment ([IQA] Boston 2012, Candela and Boston 2022), which uses rubrics to capture instruction related to mathematical tasks, explaining one's mathematical thinking and reasoning, teachers' mathematical expectations, and more. The protocol was designed for conducting live observations; however, it works equally well for video (Boston and Candela 2018). Schlesinger and colleagues (2018) also used live observations to study mathematical content, taking notice of teachers' mathematical correctness, explanations, and depth, as well as students' mathematical errors. Along those same lines, Walkington and

Marder (2018) used video observations and the UTeach Observation Protocol (UTOP) to study mathematical significance, teacher content knowledge, correctness, and more as they painted a portrait of teachers' mathematical instruction. Other researchers have used the video modality to facilitate a finer-grained analysis of mathematics teaching using the Mathematics-Scan (M-SCAN) instrument (Berry et al. 2010, Walkowiak 2014) and its various standards-based components (e.g. cognitive depth, multiple representations, mathematical connections).

Some researchers' interests pertain to particular instructional formats. Herbel-Eisenmann and Otten (2011) studied teacher and student interactions specifically during whole-class instruction. Using video recordings of mathematics classrooms, they attended to mathematical terminology, who uses those mathematical terms, and the nature of the relationships between the terms, for the purpose of determining the socially constructed mathematical meanings. This detailed discourse analysis would not have been possible through live, unrecorded observations.

Researchers have also explored student and teacher interactions during non-whole-class settings. For instance, Jansen (2012) investigated, using video recordings, teachers' actions while students worked collaboratively in small groups. Yackel and colleagues (1991) examined small-group settings but focused on student interactions rather than the teacher to understand students' mathematical activity. Using video recordings, they captured multiple small groups' interactions within a single lesson. Another way in which researchers have examined small-group interactions is by studying students' mathematical identities as they interact with their group mates (Bishop 2012). Using both live observations and video recordings, Bishop (2012) attended to both large and small-grained components of students' interactions.

As evidenced above, classroom observations, as a form of data collection, are versatile, allowing researchers to study many different aspects of classroom instruction, but the decision to do so live or by video has been a perennial question in designing the data collection process. For instance, might live observations provide a clearer picture of the lesson dynamics and lesson context than video? Or are video observations better suited for studying non-whole-class interactions compared to conducting them live? In addition to questions such as these, researchers must consider budgets, human capacity, and time (Haidet 2009, Grossman 2014). Is it more cost-effective for researchers to travel to the data collection site or to send a recording device? How much time is involved in travel compared to watching and transcribing video recordings? Does a research team have the human capacity to observe multiple participants in multiple locations? Answering such questions is non-trivial, as one's choices are driven by both the data collection logistics and the phenomenon under study. Ethical considerations must also be taken into account as schools, teachers, students, and parents may have different feelings about being observed versus being recorded, and then the security of the video data may also be of concern. From a research standpoint, it is also imperative to consider possible directional inconsistencies that might enter into the analysis due to the observation modality. Time and capacity considerations have to be weighed against the quality and the confidence in the resultant analysis and conclusions.

Framing for the present study

There are many different observation protocols used for various purposes in conducting lesson observation research in mathematics education and it is beyond the scope of one study to test the observation modalities in all instances. Thus, we selected one observation protocol with which we were familiar and which was designed to attend to multiple aspects of mathematics classroom instruction without favouring one model of instruction over another (i.e. conventional teacher-led instruction versus reform-oriented student-centred instruction). We selected the Flipped Mathematics Instruction Observation Protocol (Otten 2023a), which, despite the name, is applicable to both flipped and non-flipped lessons. It captures a variety of instructional formats at varying grain sizes and includes general quality indicators within the major components of lessons. Using this analytic framework, we pursued answers to the following research questions:

- What similarities and differences, relative to the coded aspects of instruction, emerged when comparing live lesson observations to video-recorded ones?
- What similarities and differences, relative to the coded aspects of instruction, emerged when comparing different grade-level teachers' live instruction and video recordings of their instruction?

The goal is to investigate potential differences in coding of instructional dimensions that might be related to the observation modality, rather than to the instruction itself. This is not to determine which observation modality is 'better' than the other, but rather to provide insight and awareness of how the modality might play a role in analysis and interpretation when studying particular dimensions of classroom instruction.

Method

Setting and participants

For this study, we recruited a convenience sample of K–12 teachers from the authors' professional networks who taught mathematics. The participants taught at three different schools within one midwestern state in the United States. The school pseudonyms and demographic information for each are shown in Table 1.

Within the three schools, we observed seven different teachers (grades 1–12). We observed each teacher at least once during a six-week period in Spring 2022 for a total of 18 lesson observations. Teacher demographic information is shown in Table 2.

Table 1. School demographic information.

	Race/Ethnicity							
School	Asian	Black	Hispanic	Multi Racial	Native American	White	Free or Reduced Lunch	English Language Learners
	Percentage of Students							
Southwest Urban Charter	5	27	21	6		41	53	17
Smalltown Rural Middle		3	3	5		88	24	
Smalltown Rural High			3	3		92	17	

Table 2. Participant information.

	Demographic Information				
Teacher Pseudonym	Grade (s)	School	Number of Students in Class	Number of Lessons Observed	Geographic Description
01	1	Southwest Charter	20	1	Urban
02	1	Southwest Charter	20	2	Urban
03	3	Southwest Charter	17	4	Urban
04	7	Southwest Charter	17	3	Urban
05	7	Smalltown Middle	13	2	Rural
06	8	Smalltown Middle	20	3	Rural
07	9–12	Smalltown High	23	3	Rural

Observation protocol

The observation protocol divides lessons into segments (non-instructional time, whole-class discourse, individual work time, and group work time) that indicate the purpose or format of a particular portion of the lesson. The time spent within each segment is tallied and the total time within each is recorded to provide a sense of the lesson flow and time allocation. The observation protocol also scores the lesson across four instructional clusters (Lesson Overall, Mathematical Aspects, Interactive Aspects of Whole-Class Discourse, and Interactive Aspects of Non-Whole-Class Discourse), which are further parsed into dimensions. Details for each are further elaborated in Figure 1. Within each dimension, a score of 0, 1, 2, or 3 is assigned but the score value does not necessarily mean better or worse; rather, it provides a way to distinguish what is observed. For example, when scoring Teacher Initiation during non-whole-class discourse (the teacher talking to students as they work), a score of 1 signifies the teacher did so reactively, responding to hands raised or students calling their attention. A score of 3, on the other hand, indicates the teacher proactively initiated

Observation Protocol Overview

Dimension	Description
	Lesson Overall
Focus	The extent to which an observer can identify the purpose or what is to be learned from the lesson; scores vary from 1 to 3, with 1 being, *Unclear* and 3 being, *Explicit*
Rationale	The extent to which a reason is provided for learning the lesson's focal content; scores vary from 0 to 3, with a 0 being, *No Rationale Provided* and 3 being, *Strongly Related to Content*
	Mathematical Aspects of the Lesson
Math Development	The extent to which mathematical rules, facts, and/or procedures are conceptually justified; scores vary from 0 to 3, with 0 being, *No Math Ideas* and 3 being *Conceptually Developed*
Unmitigated Math Errors	The extent to which mathematical errors are present **and** left uncorrected in the lesson; scores vary from 1 to 3, with 1 being *Several (or One Major)* and 3 being, *No Unmitigated Errors*
Math Representations	The extent to which multiple representations (e.g., symbols, figures, graphs, tables, text) are used and integrated within the lesson; scores vary from 0 to 3, with 0 being, a *Single Representation* and 3 being, *Strongly Integrated*
Lesson Connections	The extent to which the teacher connects the current lesson's material to prior or future lessons; scores vary from 0 to 3, with 0 being, *None Provided* and 3 being, *Substantial Strong*
	Interactive Aspects During Whole-Class Discourse
Student Engagement	The extent to which students are on task, doing what is expected of them (e.g., raising hands, heads off desks, attentive to the classroom activities); scores vary from 1 to 3, with 1 being, *Most Off Task Most of the Time* and 3 being, *Most on Task Most of the Time*
Students Publicly Involved	The extent to which students respond to, interact with, and participate in the classroom discourse; scores vary from 1 to 3, with 0 being, *Mostly Silent* and 3 being, *Mostly Contribute*
Student Contribution Length	When students participate in the classroom discourse, are they providing short, one-word answers or are they giving lengthier, multi-sentence explanations; scores vary from 1 to 3, with 1 being, *Low* and 3 being, *High*
Nature of Discourse	The extent to which students are attending to, building from, and/or connecting to other students' ideas; scores vary from 1 to 3, with 1 being, *Mostly Sharing* and 3 being, *Mostly Collaborative*
Math Authority	The extent to which the teacher, textbook, or classroom community determines if someone's idea is mathematically correct or decides which mathematical ideas get taken up; scores vary from 1 to 3, with 1 being, *Teacher/Textbook* and 3 being, *Class/Shared*
	Interactive Aspects of Non-Whole-Class Discourse
Student Engagement	The extent to which students are on task, doing what is expected of them (e.g., working independently, collaborating with peers, engaging with the assignment); scores vary from 1 to 3, with 1 being, *Mostly Off Task* and 3 being, *Mostly On Task*
Teacher Circulation	The extent to which the teacher is moving around the classroom; scores vary from 1 to 3, with 1 being, *Mostly Stationary* and 3 being, *Mostly Circulating*
Teacher Initiation	The ratio of interactions that are teacher initiated (i.e., proactive) compared to those that are initiated by the students (i.e., reactive); scores vary from 1 to 3, with 1 being, *Reactive* and 3 being, *Proactive*
Group Peer Talk	The extent to which students engage with one another during group work time; applies when the teacher explicitly says, "work in groups" or something similar; scores vary from 1 to 3, with 1 being, *Individual* and 3 being, *Interactive*
Independent Peer Talk	The extent to which students engage with one another during independent work time; applies when the teacher explicitly says to work independently or when it is unclear of the expectation for the work time; scores vary from 1 to 3, with 1 being, *Individual* and 3 being, *Interactive*

Figure 1. Observation protocol overview.

conversations or interactions with students, without the students calling them over. For such quantifiable codes, a 0 indicates an absence of the observable behaviour (e.g. no interactions between teacher and students during the non-whole-class discourse). Certain codes are not quantifiable; instead, the 1–3 scores simply denote different types of interaction. Such codes were Nature of Discourse and Math Authority, with 1 denoting sharing discourse and teacher/textbook authority, respectively, whereas 3 denotes collaborative discourse and shared classroom authority, and 2 indicates a mixture of both types. Again, 1 is not to be interpreted as better or worse than 3 for any code, just different. See Otten (2023a) and Otten et al. (2018) for additional details.

The protocol was designed to be employed with live observations. Two coders attend the lesson and fill out field notes in real time then use those field notes immediately after the lesson to complete the scoring on all dimensions. The coding rubric was refined until coding agreement between two coders consistently surpassed 80% (Otten et al. 2018) and, to ensure even higher reliability, the two coders would subsequently discuss the codes and form a final reconciled version of the codes that were then used as the encapsulation of the lessons. For video-recorded lessons, the protocol can still be used as long as the recording captures the publicly-displayed images or text during whole-class discourse, clear audio of the teacher and students who speak publicly and tracks the teacher during non-whole-class discourse. A Swivl robot, which holds a tablet camera and rotates, works well with the teacher wearing the tracker. As with the live process, two coders view the video-recorded lesson and subsequently form a reconciled version of the codes. Video coders have the additional benefit of being able to review a portion of the lesson video to aid in reconciling a code.

Data collection

The research team was trained to use the Flipped Mathematics Instruction Observation Protocol (Otten 2023a) over three months (January–March 2022) by two research team members who co-developed the protocol. After an initial training, team members independently coded sample videos, then met regularly to discuss, understand the rationale for protocol scores, and come to agreement. Data collection began in April and concluded in May 2022. It involved conducting lesson observations of typical mathematical lessons (i.e. not review or test days) using the Flipped Mathematics Instruction Observation Protocol (Otten 2023a). For each observation, there were live observers in the room and they had only brief interactions with the teachers, talking about the logistics of the observation and sometimes conversation about the school context but, importantly, not further detail or interpretation of the observed lesson. The live observers also video-recorded the lesson so that the video observation could be later watched and scored separately by research team members who did not participate in the live observation. The live observers, after scoring lessons independently, reconciled their coding immediately following the lesson. The video observers coded the videos independently in the subsequent days and then scheduled a time to reconcile discrepancies following their individual scoring. The reconciled scores were tabulated for each lesson, with live and video scores kept separate, and these served as the data corpus for the present study. We had 18 total observations, but during one of the observations the video was not captured due to technical difficulties. Thus, our data set included scores from 17 live observations and 17 video observations used for comparative analysis.

We used an iPad and Swivl device to video-record. The Swivl, an automated robot, used infrared technology to follow the teacher as they moved around the classroom. The teacher wore a marker that captured the audio and determined the field of vision for the Swivl robot. An iPad rested on the Swivl robot and served as the video recording device (shown in Figure 2).

The video recording was stored locally (i.e. not a live stream) on the iPad and automatically uploaded to swivl.com via the Swivl app after the lesson. As the teacher moved around the classroom, the video and Swivl robot jointly turned to follow, the result being a video in which the teacher was the central focus. This was both a strength and limitation of the Swivl as it captured the teacher's actions and audio well, but was unable to capture the classroom in its entirety (i.e.

Figure 2. An image of the iPad and Swivl robot set-up.

students were at times out of view). Additionally, the Swivl allowed us to record the classroom lesson without a person physically operating the camera (although, for this study, we did have research team members present in the classrooms, we were interested in testing the feasibility of the Swivl video recording process for the future when team members would not necessarily be present). Others with different camera setups (e.g. a single stationary camera, omnidirectional camera) would likely have different results.

Data analysis

To answer the research questions, we compared both the live and video scores for each dimension in the observation protocol to determine the extent to which live and video coders agreed on a dimension's score. We noted patterns of directional inconsistency, that is, when the live or video

observation consistently scored higher across lessons. We use the term 'inconsistency' to denote the directionality of the mismatch in scores between live and video observers. We are not claiming to have evidence of 'true' scores nor is 'inconsistency' a deviation from a 'correct' score. Instead, we consider any given live or video score to be an accurate representation of what was observed within the modality with its inherent affordances and limitations. As an example, imagine scoring the Rationale dimension of the rubric. If the live coders scored the lesson rationale as a 2 (the teacher briefly stated the importance of the lesson in relation to mathematical ideas) and the video coders scored it as a 1 (the teacher briefly stated the importance of the lesson but it was unrelated to mathematical ideas), then we would note that there was inconsistency between the modalities and that live scored higher. The reason for the inconsistency might be the live coders noticed something the video coders did not, or to those in-person the rationale as to why a topic was important was more convincing than when observed on video. Neither should be interpreted as the 'true' or 'correct' score.

We repeated this comparative process of searching for consistency and inconsistency for each dimension across each of the 17 lessons. We then calculated the percent consistency, inconsistency with live higher, and inconsistency with video higher for each dimension for all 17 lessons. To address our first research question, we report observational consistency and inconsistency in relation to our four code clusters (as shown in Figure 1). Within each cluster, we report on each dimension in order from highest agreement to lowest. To address our second research question, we used the same calculations but examined similarities and differences between the seven teachers. We note relatively high or low levels of consistency – reported as percentages – and then examine two teacher cases more closely, one in which a teacher consistently scored higher when observed live (Teacher 05) and one who consistently scored higher when observed on video (Teacher 02). We elaborate qualitatively on each case to provide a fuller sense of the teachers' instruction. For clarity, after each set of findings we briefly discuss potential explanations for the consistencies and inconsistencies. We conclude with a broad discussion of considerations when selecting an observation modality.

Findings

Considering observational consistency and directional inconsistency for code clusters

This section focuses on observation codes and clusters of codes, examining whether live observations and video observations yielded consistent or inconsistent results for those codes out of the 17 lesson observations. There are four clusters in the observation protocol and we discuss possible interpretations for each in turn.

1a. Lesson overall findings

Within this cluster, there were two dimensions, one being the extent to which an observer could identify the lesson topic or purpose (Focus), and the second was the extent to which a reason was provided for why students should learn the lesson (Rationale). For Focus, there was general consistency between video and live coding (71%); however, when there was a directional inconsistency between the two modalities, live coders always scored the dimension higher than video coders (29%). For Rationale, the consistency between the two modalities was not particularly high (53%). When there was inconsistency, neither modality seemed to systematically score higher than the other (Figure 3).

1b. Lesson overall discussion

We found a systematic directional inconsistency with regard to the Lesson Focus; all inconsistencies between modalities favoured a higher live score. One potential explanation is the observer being physically present afforded opportunities for them to notice aspects of the lesson that exist on the periphery (e.g. ambient noise, student conversations and subtle movements), allowing for a better overall sense of the room (Casabianca et al. 2013, Curby et al. 2016). Although live coders

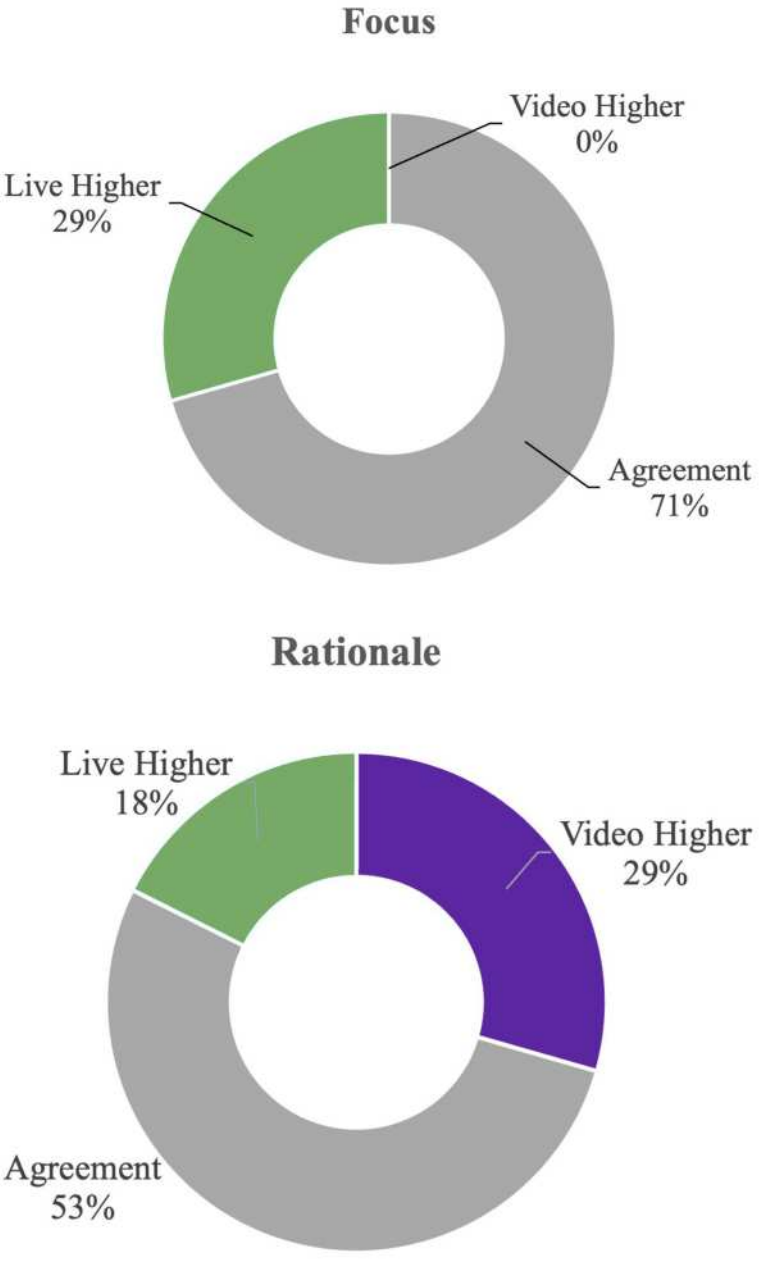

Figure 3. Graphs for Focus and Rationale consistency and inconsistency, as percentages.

were supposed to code the Focus dimension based on explicit descriptions of the lesson, they may have gotten a more implicit sense of the Focus, causing the live scores to be higher. Video coders may have not been privy to the more implicit aspects and, consequently, only marked the Focus if there was a clear, explicit description of the topic or objective of the lesson.

Another plausible explanation for the systematic directional inconsistency towards live observations for this dimension might be a function of the observation tool itself. A lesson could score a 3 for Focus if the purpose of the lesson was clear and either stated verbally or written on the board. A limitation of the Swivl was, at times, it was difficult to discern what was written on the

board due to the Swivl's proximity to the board, line of sight, or room lighting. If the teacher did not verbally state the lesson focus, having it only written down, live observers would have made notes whereas those coding the video would not have that opportunity.

The Rationale dimension did not have high levels of agreement but there was also no systematic directional inconsistency we could detect toward either of the modalities. This finding suggests general caution with respect to the reliability of this code but the modalities of observation seem to be on equal footing.

2a. Mathematical aspects of the lesson findings

The Mathematical Aspects of the Lesson cluster consisted of four dimensions involving the extent to which a teacher connected the current lesson to prior or future ones (Lesson Connections), mathematical correctness of explanations and solutions (Unmitigated Math Errors), integration of multiple representations (Math Representations), and justification of concepts and procedures (Math Development). There was moderate agreement between live and video modalities for three dimensions (Lesson Connections, Unmitigated Math Errors, Math Representations) and the disagreements did not tend to systematically favour live or video. Math Development was the exception as it had a low agreement (29.4%) and a pronounced directional inconsistency with 83% of the disagreements yielding a higher live coder score than on video (Figure 4).

2b. Mathematical aspects of the lesson discussion

One of the more compelling reasons why the Math Development dimension scored higher live than on video may pertain to what was being measured. The Math Development dimension required coders to cumulatively attend to the ways in which the teacher conceptually justified procedures

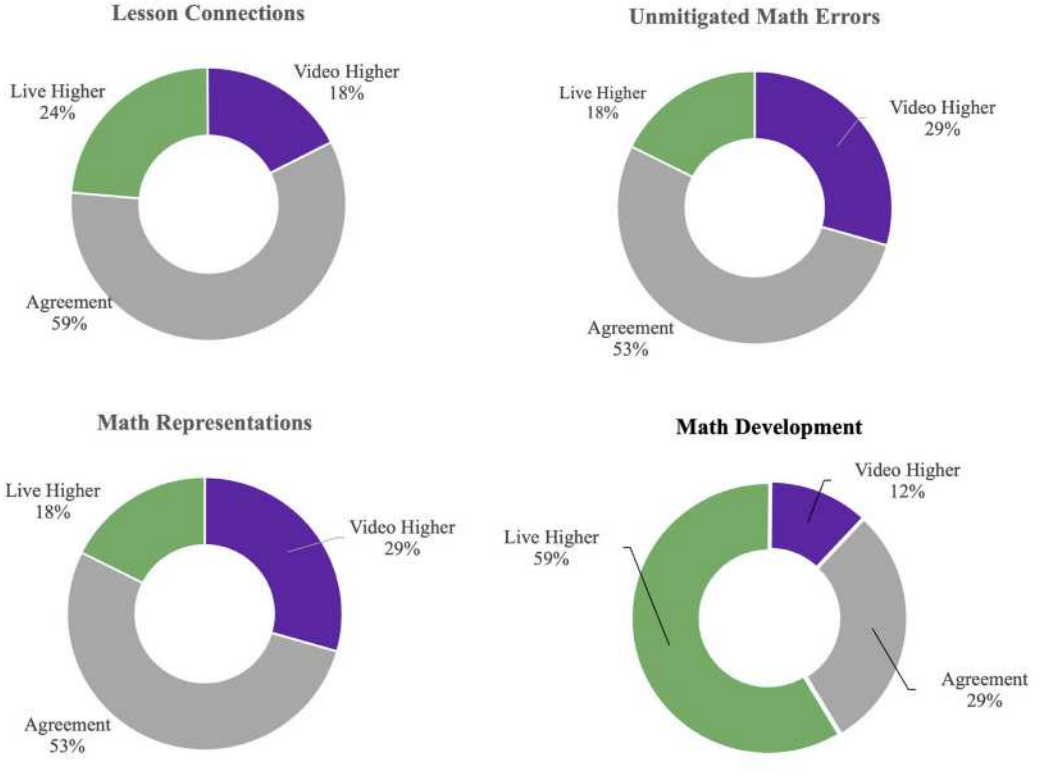

Figure 4. Graphs for lesson Connections, Unmitigated Math Errors, Math Representations, Math Development consistency and inconsistency, as percentages.

and mathematical ideas. This often required a coder to note how ideas built across a lesson, rather than bounding them to a single point in time as happened readily with the other three dimensions. Jaeger (1993) noted live observers can only attend to so much during an observation, which made us wonder, were our video coders more stringent because they had the ability to pause, rewind, and replay? Researchers have noted such features of video as a strength of the modality (Otrel-Cass et al. 2010) allowing coders to attend to finer-grained details of a lesson that may potentially be overlooked when observing live (Erickson 2006, Goldman and McDermott 2007, Lemke 2007, Jewitt 2012, Grossman 2014, Blikstad-Balas 2017). Were video coders able to attend more closely to the exact words or explanations given? Was it the case that live coders attended to the big picture rather than the exact wording because they were only able to hear it once? We hypothesise live coders had a more holistic sense of the lesson as they were inside the lesson as it was happening (e.g. student facial expressions, 'temperature' of the room). They may not have been able to capture the exact wordings; however, they could potentially observe more nonverbal actions or cues, especially of students, and video coders did not have as much access to video of students until it tracked the teacher working with small groups of students. These are questions worth considering when deciding on a modality for conducting observations. People might assume mathematical content is 'objective' and will be discerned the same on video or live, but we did not find high consistency between the two modalities.

3a. Interactive aspects of the whole-class discourse findings

This cluster included five dimensions related to whole-class discourse involving the teacher and students. These dimensions addressed the ways in which students engaged with one another's ideas (Nature of Discourse), who held the mathematical authority (Math Authority), the extent to which students were on task (Engagement), the length of students' public contributions (Contribution Length), and how many students contributed to the public discourse (Students Publicly Involved). This cluster was characterised by generally high agreement between the two modalities with four of the five dimensions achieving at least 88% agreement and two dimensions yielding 100% agreement. The noteworthy inconsistency within this cluster was when coders scored the Students Publicly Involved dimension, which had only 64.7% agreement between live and video. Moreover, when there was inconsistency between the two modalities (35.3% of lessons), live coders always scored Students Publicly Involved higher than video coders (Figure 5).

3b. Interactive aspects of whole-class discourse discussion

Contrary to the mathematical aspects cluster, there was high consistency around the interactions within whole-class discourse. The high consistency is largely explained by the lack of variation in these dimensions generally. In our observations here and elsewhere (Otten 2023b), the mathematical authority tends to be the teacher, and students tend to engage in discourse with short turns characterised by sharing their own thoughts (i.e. sharing) rather than building from their peers' ideas (i.e. collaborative). The dimension Students Publicly Involved was the exception, with all inconsistencies favouring a higher live coder score. This dimension was assessed by noting the proportion of students in the class who responded, interacted with, and publicly participated in the classroom discourse. The directional inconsistency toward live coders scoring this dimension higher was not surprising and can most likely be attributed to the camera's field of vision. A commonly-cited limitation of video is the restricted view of the camera's field of vision (Jaeger 1993, Holm 2008, Casabianca et al. 2013, Curby et al. 2016, Blikstad-Balas 2017) and is dictated by the videographer (Jewitt 2012, Ing and Samkian 2018). Our camera tracked the teacher, not the students, so it was difficult at times for the video coders to view exactly which students were speaking and how many. Others have described this disadvantage of video as an affordance of being live and have suggested live observations allow the observer to scan the entire room and notice peripheral interactions a video may miss (Casabianca et al. 2013, Curby et al. 2016).

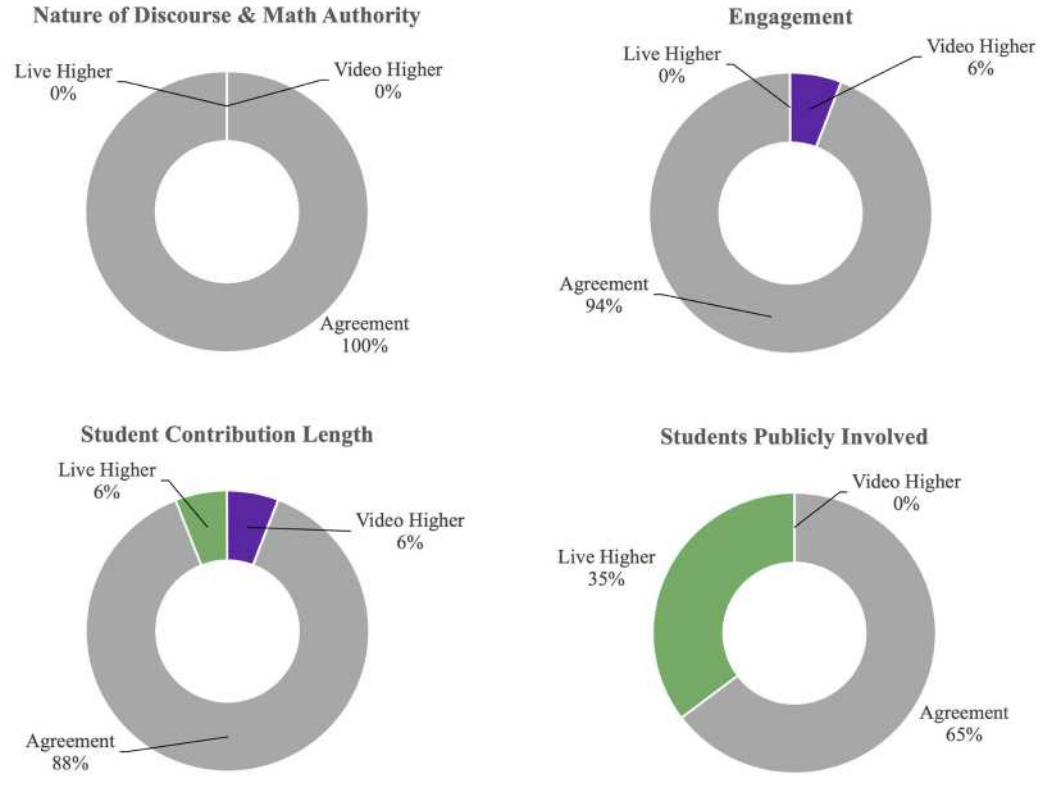

Figure 5. Graphs for Interactive Aspects of Whole-Class Discourse consistency and inconsistency, as percentages.

4a. Interactive aspects of the non-whole-class discourse findings

The last code cluster pertained to the student and teacher interactions during non-whole-class discourse (i.e. individual work time, group work time). For Teacher Circulation (i.e. the extent to which the teacher moved around the classroom) and Student Engagement (i.e. the extent to which students were observably on task), there was high agreement between the live and video scores, 94.1% and 88.2% respectively. For group Peer Talk and independent Peer Talk (i.e. the extent to which students interact with each other during the time designated for working in groups or individually) there was moderate agreement (76.5% and 70.6% respectively). There was a slight directional inconsistency toward video coders scoring the group Peer Talk higher (17.6%) and live coders scoring independent Peer Talk higher (23.5%). For Teacher Initiation (i.e. the extent to which the teacher interacted with students reactively or proactively), there was mild agreement (53%) but no clear directional inconsistency between the two modalities (Figure 6).

4b. Interactive aspects of the non-whole-class discourse discussion

Within this last cluster, there were some interesting findings. The Student Engagement dimension had a high agreement (88%). Prior to data collection, we expected substantial differences between the live and video scores for Student Engagement because we thought live coders might have a different sense of the room and students being on task. This finding of agreement offers an alternative perspective from prior research which suggests live coders may be able to pick up on behaviours on the periphery, outside the camera's field of vision (Casabianca et al. 2013, Curby et al. 2016).

There was inconsistency around the two Peer Talk dimensions, group and independent. As a reminder, the group format refers to a segment of the lesson where the students are explicitly

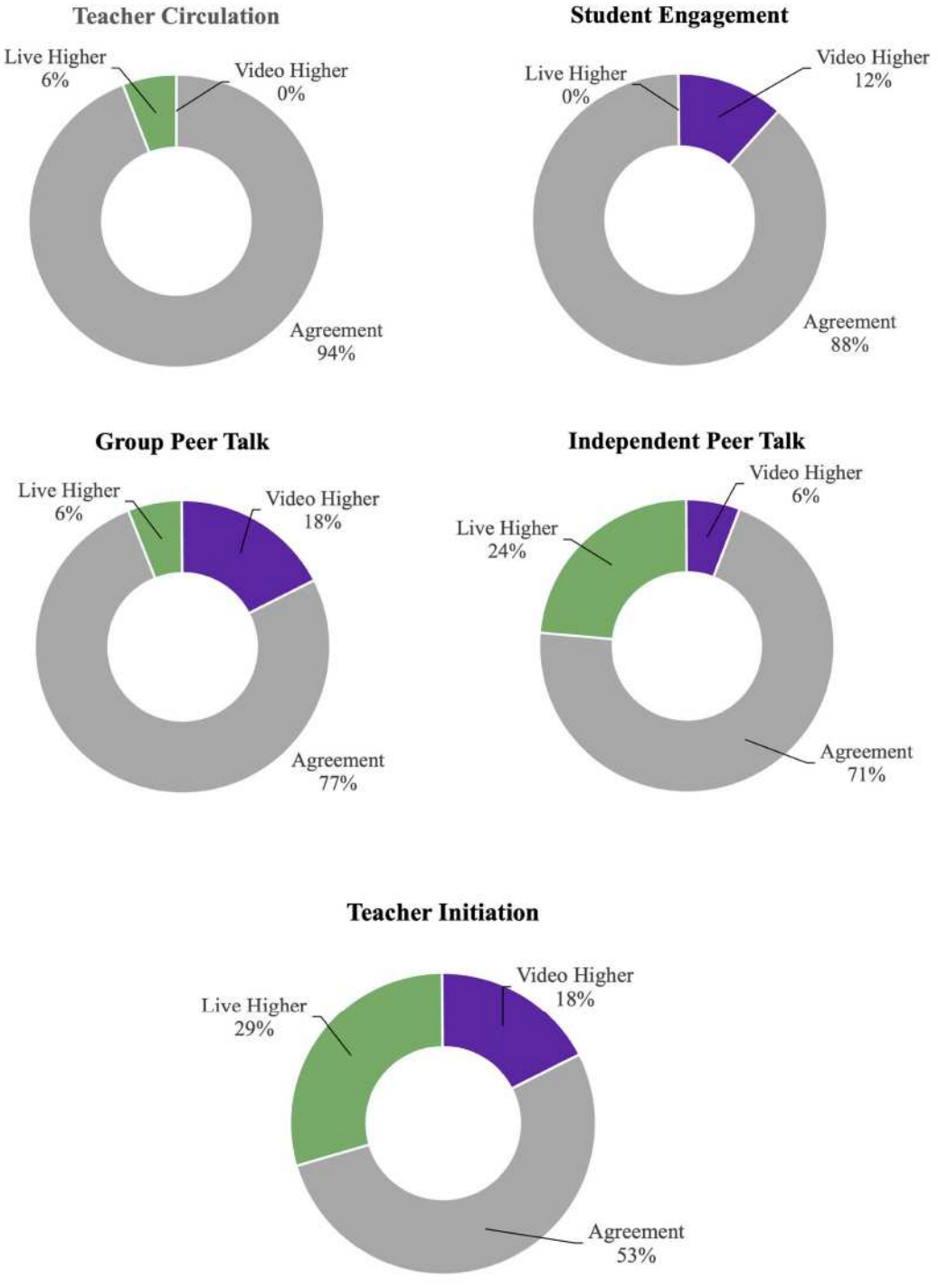

Figure 6. Graphs of Interactive Aspects of the Non-Whole-Class Discourse, as percentages.

expected to work in groups whereas the independent format is when students are allowed to work alone (though some may freely choose to collaborate). Peer Talk, in either format, refers to the actual level of student-to-student talk during these work times. Video coders tended to score group Peer

Talk higher (18%), while live coders tended to do so for independent Peer Talk (24%). One possible explanation is peer talk was more evident on the video during group work time because of classroom norms for group work. Although we are not certain of the exact classroom norms, it seemed as though teachers expected their students to talk to their peers and they did so with more volume during group work than during independent work time, thus being picked up more clearly on video. In contrast, it may have been that students were more conscious of their volume during independent work because they were unsure of the teacher's expectation with regard to allowing students to collaborate, or students were trying to be considerate of peers who were trying to work quietly alone. As such, it was reasonable for live scores to be higher for the independent Peer Talk dimension because live coders had a fuller sense of the room and may have been able to pick up on quiet conversations that the video recording missed since students were not wearing microphones.

The Teacher initiation dimension had the least agreement between the modalities, with neither appearing to have a systematic directional inconsistency. Our camera setup tracked the teacher, which made it difficult to notice if and when an off-screen student raised their hand or signalled to the teacher to come over. When this happened, on video it appeared as though the teacher was being proactive rather than reactive. This phenomenon is consistent with prior research about actions occurring off camera (e.g. Casabianca et al. 2013). For the live coders in our study, it was challenging at times to discern what a teacher said – live coders were not privy to the microphone audio. There was also the potential to miss a student getting the teacher's attention because in many instances the students made small, fleeting gestures for the teacher's attention. Although we cannot be certain about the reason for the differences, prior research has recognised these limitations of live coding as affordances of video (e.g. putting a microphone on the teacher, attending to specific moments in a video) (Otrel-Cass et al. 2010, Casabianca et al. 2013, Curby et al. 2016). Thus, even though both modalities of coder are able to watch the teacher circulating during student work, each modality has its limitations with regard to discerning the precise initiation of the teacher-student interactions.

Considering observational consistency and directional inconsistency for teachers

In this section, we consider whether a particular teacher's instruction was captured in consistent or inconsistent ways across the two modalities of observation. Perhaps the overall facets and nuances of a teacher's instruction may show up differently in a video recording than it appeared to live observers. When looking at the comparison between video and live scores for individual teachers, there was between 73 and 75% consistency for five of the seven teachers (Table 3). The other two teachers (02 and 05) had consistency on 63% or 66% of the codes.

We did not do a sufficient number of observations for each teacher to fully characterise their teaching practice, as this was not the purpose of the study, and so we cannot fully explore how the consistencies or inconsistencies relate to their general teaching practice. Additionally, because of our limited teacher sample at various grade levels and in various school contexts, we cannot draw conclusions about a potential directional inconsistency toward higher video scores in lower grade levels or in urban settings, nor about a potential directional inconsistency toward higher

Table 3. Percent agreement by teacher.

	Teacher						
	01	02	03	04	05	06	07
Agreement	75	63	73	75	66	75	73
Live Higher	19	9	8	13	31	23	19
Video Higher	6	28	19	13	3	2	8

Note. All percentages are rounded to the nearest percent.

Table 4. Percent agreement for teacher 02 and 05 for each dimension.

Cluster	Dimension	Percent Agreement	
		Teacher 02	Teacher 05
Lesson Overall	Focus	100	0
	Rationale	0	50
Mathematical Aspects of the Lesson	Math Development	50	0
	Unmitigated Math Errors	50	100
	Math Representations	0	50
	Lesson Connections	50	100
Interactive Aspects of Whole-Class Discourse	Student Engagement	100	100
	Students Publicly Involved	50	100
	Student Contribution Length	100	50
	Nature of Discourse	100	100
	Math Authority	100	100
Interactive Aspects of Non-Whole-Class Discourse	Student Engagement	0	100
	Teacher Circulation	100	50
	Teacher Initiation	50	50
	Group Peer Talk	50	50
	Independent Peer Talk	100	50

Note. Shaded cells (striped or solid) denote dimensions where there was 0% or 50% agreement between live and video. Striped cells denote a higher live score and those fully shaded denote a higher video score.

live scores in higher grade levels or in rural settings. We can, however, examine more closely the two most extreme cases, Teachers 02 and 05, to get a sense of two specific examples and how a directional inconsistency toward video or live coding may have occurred in our analysis (Table 4).

A case of video codes higher, Teacher 02

Recall Teacher 02 is in a grade 1 classroom in an urban school. Our research team observed Teacher 02 live and again on video on two different occasions. During the lessons, Teacher 02 taught her students how to add and subtract within the context of story problems (lesson 1) and used place-value strategies to find the sum (e.g. 28 + 34) of two numbers (lesson 2). During instruction, Teacher 02 alternated between whole-class discussion and independent/group work about ten times (e.g. whole-class, non-whole-class, whole-class) splitting the class time evenly between the two formats. As they worked, students used a variety of solution strategies (including using manipulatives) and the teacher frequently focused her questions on those strategies. The seating arrangement involved students sitting on the floor at the front of the room or at desks; students appeared to have their choice of seating.

For Teacher 02, there were two findings of particular interest. First, the live and video coders typically agreed on the Interactive Aspects of Whole-Class Discourse, which suggests coders were able to get the same sense of the room regardless of modality. Second, the Mathematical Aspects of the Lesson were coded more favourably on video than by live coders. Although it was unclear as to what caused the disagreements, potential explanations might be related to the classroom environment. During whole-class discourse segments of the lesson, students used manipulatives and the noise level in the room was high. It may have been the case that live observers were unable to hear the teacher's explanation as clearly as the video coders. Similarly, this was a first-grade classroom where students moved around the room frequently, with some students sitting at desks while others were on the floor. The Swivl may have been better able to observe the teacher as she moved between students, possibly making mathematical points or drawing connections. Alternatively, the live coders sat on the opposite side of the room from the Swivl and thus, they had a different perspective of the room. For example, the students sitting on the floor were closer to the Swivl than they were to the live observers, and in the commotion, it is possible live coders missed some aspects of mathematical development. Again, this is not to say the video coding was 'correct', because if live coders missed some mathematical development, it is plausible students

missed it as well, and thus it could be argued the lower live scores were more representative of the learning opportunities in the lesson.

A case of live codes higher, Teacher 05

Teacher 05 taught 7th grade in a rural setting. During the first observation, Teacher 05 alternated between whole-class and non-whole-class discourse frequently, spending a minute or two in one format before changing to another. Her lesson was focused on proportions, ratios, and percents embedded within the context of student attendance at a school dance. The teacher would spend a minute or two introducing a problem, let the students work in groups or individually, and then bring the class back together for the next problem. For the second observation, Teacher 05 had students working on an in-class project in which students were creating a food truck menu and applying their knowledge of percents, discounts, and tax. There were fewer changes from whole-class to non-whole-class formats than in the first observation, with over half the class time spent working individually. For both lessons, the students sat at individual desks, arranged in rows, with two to four students per row.

Across the two observations, the video and live coders agreed on several dimensions of Teacher 05's instruction, especially regarding the Interactive Aspects of Whole-Class Discourse (Figure 1). Many of the other dimensions, however, were scored higher by live coders than by the video coders. For example, the live coders viewed Teacher 05's Mathematical Development as being more substantial than the video coders perceived it. Similarly, there was little agreement regarding the Lesson Overall cluster. The protocol is such that for Lesson Focus, teachers receive a score of 2 or 3 if the focus is inferable or explicit, respectively, and explicit mentions can be done verbally or visually. One potential (and likely) explanation for the disagreement with Focus was it was written on the board and video coders were not privy to it in the video. A teacher receives a score of 1 if the Focus is unclear (i.e. not discernible to the observer). Thus, the directional inconsistency toward scoring the Lesson Focus higher live than on video was probably a function of the observation tool rather than the observation modality.

Other inconsistencies in the coding of Teacher 05's lessons related to interactions. The live coders marked Student Contribution Length as moderate for both lessons (video coders thought the student contributions in the first lesson were predominantly short) and, during the non-whole-class discourse, the live coders judged the Student Peer Talk to be high (video coders tended to view it as moderate). One of the more convincing explanations for the difference in modality scores was the Swivl itself. As previously mentioned, the Swivl tracked and the microphone was attached to the teacher. It may have been the case that the video was unable to capture the student interactions to the same extent as the live coders discerned.

There was one instance where the video coders scored a dimension higher than the live coders – Teacher Initiation – meaning the video coders considered Teacher 05's interactions with students to be more proactive than reactive. This was not surprising as students often made small gestures to signal they needed assistance rather than more pronounced gestures (e.g. raising one's hand high in the air). On video, it appeared as though the teacher was proactive as she interacted with students, but the live scores suggest they considered the interactions to be more reactive.

Conclusion

The methodological decision to conduct video or live classroom observations has been long contemplated (Jaeger 1993). There are affordances and constraints to both modalities that researchers need to carefully consider. Prior research has noted differences exist between the modalities and video and live classroom observations are not interchangeable (Casabianca et al. 2013, Curby et al. 2016, Gridley et al. 2018). We also found differences and, moreover, noticed there seems to be a directional inconsistency of live or video scoring depending on the phenomena under study. For example, when determining the extent to which students are publicly involved during whole-

class discourse in our study, live coders always scored this dimension higher when there was inconsistency between modalities. This is noteworthy for researchers studying student interactions as they decide on a modality of data collection, but also something to consider during data analysis because video observations may be artificially low or live ones may be artificially high.

Another set of dimensions to carefully consider are those within the Mathematical Aspects of the Lesson cluster. Agreement between the two modalities was mild with no clear directional inconsistency towards video or live for three of the four dimensions. We have offered potential explanations for the directional inconsistencies, but more importantly, we would encourage researchers collecting observational data to be aware of possible dynamics in scoring due to the chosen modality. These dimensions are more difficult to score because they are focused on the interaction's content and meaning rather than the interactions themselves. They often are not contained to a single-moment or interaction in time, requiring an observer to keep detailed field notes of pivotal moments as ideas build.

In contrast, there were a number of dimensions that had high agreement between video and live observation. These dimensions attended to more observable (audio or visual) aspects of classroom interactions (e.g. Student Engagement, Teacher Circulation, Student Contribution Length) or were dimensions that did not vary (Nature of Discourse, Math Authority) and these may have ultimately been easier to capture regardless of modality. Exceptions to agreement of these more observable dimensions are Teacher Initiation and Students Publicly Involved, with live observers tending to score these two dimensions higher than video observers.

Although we have shed light on some nuances with regard to these dimensions of mathematics instruction and the modalities of observation, placing our findings into dialogue with past research on this area, it is important to reiterate the limitations of this study. We did not conduct enough observations to saturate our understanding of the teachers' instructional styles, but the study did allow us to examine possible inconsistencies that can occur within specific lesson occurrences such as students' public involvement during a whole-class interaction or teacher circulation during work time. The phenomena that arose in such segments were of immediate interest, even without large numbers of observations. Nevertheless, the findings should be taken as suggestive of possible coding patterns, not definitive. Furthermore, the findings may be unique to mathematics instruction (though we think some findings, such as those involving the interactional, not mathematical, dimensions may be reasonably hypothesised as extending to other subject areas) and are most certainly limited to the dimensions as defined by Observation Protocol B. Other definitions or operationalisations of dimensions of instruction could have different patterns in live versus video coding.

Through our data collection and analysis, we did not identify a 'true' score and the extent to which video or live coders deviated from it. Instead, our purpose was to identify inconsistencies between the two modalities and note dimensions in which there seemed to be systematic directional inconsistency. It may well be the case that either one is artificially high or low, but that is beyond the scope of this analysis and may be a site for future research. Despite this limitation, we do contend there are some dimensions of instruction that appear to be unaffected by modality and others that are. As researchers make decisions about how best to collect data, topics of conversation often include human capacity (Lemke 2007, Jewitt 2012, Blikstad-Balas 2017, Gridley et al. 2018), cost of travel (Jaeger 1993, Grossman 2014, Curby et al. 2016), availability of technology, or the ability to create a permanent record (Jacobs et al. 2007, Holm 2008, Derry et al. 2010, Otrel-Cass et al. 2010, Casabianca et al. 2013). In addition to these considerations, we would encourage researchers to thoughtfully consider the focus of their inquiry and weigh the modality's affordances and constraints as it pertains to that focus.

Acknowledgments

This work was supported by the National Science Foundation under Grant 2206774 (de Araujo, PI), though any opinions, findings, or conclusions expressed here are those of the authors and do not necessarily reflect the views of the NSF. The

authors thank Maria Stewart, Courtney Vahle, Mitchelle Wambua, and Faustina Baah for helping with data collection. The authors also thank the teachers and students for letting us observe them and learn from their interactions.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by National Science Foundation: [Grant Number 2206774].

ORCID

F. Paul Wonsavage http://orcid.org/0000-0003-1486-3790
Samuel Otten http://orcid.org/0000-0002-3496-2078
Amber G. Candela http://orcid.org/0000-0003-2920-2397
Zandra de Araujo http://orcid.org/0000-0002-8186-6599

References

Adamson, S.L., et al., 2003. Reformed undergraduate instruction and Its subsequent impact on secondary school teaching practice and student achievement. *Journal of research in science teaching*, 40 (10), 939–957. doi:10.1002/tea.10117.

Amrein-Beardsley, A., Popp, O., and Sharon, E., 2012. Peer observations Among faculty in a college of education: investigating the summative and formative uses of the reformed teaching observation protocol (RTOP). *Educational assessment, evaluation and accountability*, 24, 5–24. doi:10.1007/s11092-011-9135-1.

Ayuwanti, I., Marsigit, M., and Siswoyo, D., 2021. Teacher-Student interaction in mathematics learning. *International journal of evaluation and research in education*, 10 (2), 660–667. doi:10.11591/ijere.v10i2.21184.

Berry, I.I.I., et al. 2010. The mathematics scan (M-scan): A measure of mathematics instructional quality. Unpublished measure. University of Virginia.

Bishop, J.P., 2012. She's always been the smart One. I've always been the dumb one': identities in the mathematics classroom. *Journal for research in mathematics education*, 43 (1), 34–74. doi:10.5951/jresematheduc.43.1.0034.

Blikstad-Balas, M., 2017. Key challenges of using video when investigating social practices in education: contextualization, magnification, and representation. *International journal of research & method in education*, 40 (5), 511–523. doi:10.1080/1743727X.2016.1181162.

Borko, H., et al., 2005. Artifact packages for characterizing classroom practice: A pilot study. *Educational assessment*, 10 (2), 73–104. doi:10.1207/s15326977ea1002_1.

Bostic, J., et al., 2021. Classroom observation and mathematics education research. *Journal of mathematics teacher education*, 24 (1), 5–31. doi:10.1007/s10857-019-09445-0.

Boston, M., 2012. Assessing instructional quality in mathematics. *The elementary school journal*, 113 (1), 76–104. doi:10.1086/666387.

Boston, M., et al., 2015. A comparison of mathematics classroom observation protocols. *Mathematics teacher educator*, 3, 154–175. doi:10.5951/mathteaceduc.3.2.0154.

Boston, M.D., and Candela, A.G., 2018. The instructional quality assessment as a tool for reflecting on instructional practice. *ZDM*, 50, 427–444. doi:10.1007/s11858-018-0916-6.

Candela, A.G., and Boston, M., 2022. Centering professional development around the instructional quality assessment rubrics. *Mathematics teacher educator*, 10 (3), 204–222. doi:10.5951/MTE.2021.0013.

Casabianca, J.M., et al., 2013. Effect of observation mode on measures of secondary mathematics teaching. *Educational and psychological measurement*, 73 (5), 757–783. doi:10.1177/0013164413486987.

Chilton, H., and McCracken, W., 2017. New technology, changing pedagogies? exploring the concept of remote teaching placement supervision. *Higher education pedagogies*, 2 (1), 116–130. doi:10.1080/23752696.2017.1366276.

Curby, T.W., et al., 2016. Live versus video observations: comparing the reliability and validity of Two methods of assessing classroom quality. *Journal of psychoeducational assessment*, 34 (8), 765–781. doi:10.1177/0734282915627115.

Derry, S.J., et al., 2010. Conducting video research in the learning sciences: guidance on selection, analysis, technology, and ethics. *The journal of the learning sciences*, 19 (1), 3–53. doi:10.1080/10508400903452884.

Dockerty, K., 2022. Training teachers during the COVID-19 pandemic: using live video for observation of practicum. *Research on education and media*, 14 (2), 15–21. doi:10.2478/rem-2022-0017.

Erickson, F., 2006. Definition and analysis of data from videotape: some research procedures and their rationales, edited by J. L. Green, G. Camilli, and P. B. Elmore, eds. *Handbook of complementary methods in education research*. New York: Routledge, 177–191.

Ferdig, R.E., and Kosko, K.W., 2020. Implementing 360 video to increase immersion, perceptual capacity, and teacher noticing. *Techtrends*, 64 (6), 849–859. doi:10.1007/s11528-020-00522-3.

Gold, B., and Windscheid, J., 2020. Observing 360-degree classroom videos–effects of video type on presence, emotions, workload, classroom observations, and ratings of teaching quality. *Computers and education*, 156, 103960. doi:10.1016/j.compedu.2020.103960.

Goldman, S., and McDermott, R., 2007. Staying the course with video analysis, edited by R. Goldman, R. Pea, B. Barron, and S. J. Derry, eds. *Video research in the learning sciences*. New York: Routledge, 101–114.

Gridley, N., Bywater, T.J., and Hutchings, J.M., 2018. Comparing live and video observation to assess early parent-child interactions in the home. *Journal of child and family studies*, 27 (6), 1818–1829. doi:10.1007/s10826-018-1039-y.

Grossman, P., 2014, November. Collecting evidence of instruction with video and observation data in NCES surveys. *Prepared for the national academy of education's workshop to examine current and potential uses of NCES longitudinal surveys by the education research community*.

Grouws, D.A., et al., 2010. Mathematics teachers' use of instructional time and relationships to textbook content organization and class period format. In: *Hawaii international conference on education*. Honolulu, HI, 1–15.

Haidet, K.K., et al., 2009. Methods to improve reliability of video-recorded behavioral data. *Research in nursing and health*, 32 (4), 465–474. doi:10.1002/nur.20334.

Herbel-Eisenmann, B.A., and Otten, S., 2011. Mapping mathematics in classroom discourse. *Journal for research in mathematics education*, 42 (5), 451–485. doi:10.5951/jresematheduc.42.5.0451.

Holm, G., 2008. Visual research methods: where are we and where are we going. In: S. N. Hesse-Biber, and P. Leavy, eds. *Handbook of emergent methods*. City: Guilford Press, 325–341.

Ing, M., and Samkian, A., 2018. Research commentary: raising concerns about sharing and reusing large-scale mathematics classroom observation video data. *Journal for research in mathematics education*, 49 (3), 247–260. doi:10.5951/jresematheduc.49.3.0247.

Ing, M., and Webb, N.M., 2012. Characterizing mathematics classroom practice: impact of observation and coding choices. *Educational measurement: issues and practice*, 31, 14–26. doi:10.1111/j.1745-3992.2011.00224.x.

Jacobs, J.K., Hollingsworth, H., and Givvin, K.B., 2007. Video-based research made 'easy': methodological lessons learned from the TIMSS video studies. *Field methods*, 19 (3), 284–299. doi:10.1177/1525822X07302106.

Jaeger, R.M. 1993, April. *Live vs. memorex: psychometric and practical issues in the collection of data on teachers' performances in the classroom*. Paper presented at the *annual meeting of the American educational research association*. Atlanta, GA.

Jansen, A., 2012. Developing productive dispositions during small-group work in Two sixth grade mathematics classrooms. *Middle grades research journal*, 7 (1), 37–56.

Jewitt, C., 2012. *An introduction to using video for research*. London: National Centre for Research Methods, 1–25.

Lemke, J., 2007. Video epistemology in-and-outside the box: traversing attentional spaces. In: R. Goldman, R. Pea, B. Barron, S. J. Derry, eds. *Video research in the learning sciences*. New York, NY: Routledge, 39–51.

Mac Mahon, B., Ó Grádaigh, S., and Ní Ghuidhir, S., 2019. Super vision: The role of remote observation in the professional learning of student teachers and novice placement tutors. *Techtrends*, 63 (6), 703–710. doi:10.1007/s11528-019-00432-z.

Otrel-Cass, K., Cowie, B., and Maguire, M., 2010. Taking video cameras into the classroom. *Waikato journal of education*, 15 (2), 109–118. doi:10.15663/wje.v15i2.117.

Otten, S., et al., 2023a. A framework for capturing structural variation in flipped mathematics instruction. *International journal of mathematical education in science and technology*, 54 (5), 639–670.

Otten, S., et al., 2023b. When whole-class discourse predicts poor learning outcomes: An examination of 47 secondary algebra classes. In: T. Lamberg, ed. *Proceedings of the 45th annual meeting of the north American chapter of the international group for the psychology of mathematics education*. Reno, NV: PME-NA, 1007–1011.

Otten, S., de Araujo, Z., and Sherman, M., 2018. Capturing variability in flipped mathematics instruction. In: T. E. Hodges, G. J. Roy, and A. M. Tyminski, eds. *Proceedings of the 40th annual meeting of the north American chapter of the international group for the psychology of mathematics education*. Greenville, SC: Clemson University and University of South Carolina, 1052–1059.

Ramakrishnan, A., et al., 2023. Toward automated classroom observation: multimodal machine learning to estimate CLASS positive climate and negative climate. *Ieee transactions on affective computing*, 14 (1), 664–679. doi:10.1109/TAFFC.2021.3059209.

Ryan, A.M., et al., 1995. Direct, indirect, and controlled observation and rating accuracy. *Journal of applied psychology*, 80 (6), 664–670. doi:10.1037/0021-9010.80.6.664.

Sawada, D., et al., 2002. Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School science and mathematics*, 102, 245–253. doi:10.1111/j.1949-8594.2002.tb17883.x.

Schlesinger, L., et al., 2018. Subject-Specific characteristics of instructional quality in mathematics education. *ZDM*, 50, 475–490. doi:10.1007/s11858-018-0917-5.

Star, J.R., and Strickland, S.K., 2008. Learning to observe: using video to improve preservice mathematics teachers' ability to notice. *Journal of mathematics teacher education*, 11, 107–125. doi:10.1007/s10857-007-9063-7.

Walkington, C., and Marder, M., 2018. Using the UTeach observation protocol (UTOP) to understand the quality of mathematics instruction. *ZDM*, 50, 507–519. doi:10.1007/s11858-018-0923-7.

Walkowiak, T.A., et al., 2014. Introducing an observational measure of standards-based mathematics teaching practices: evidence of validity and score reliability. *Educational studies in mathematics*, 85, 109–128. doi:10.1007/s10649-013-9499-x.

Weiss, I.R., et al., 2003. *Looking inside the classroom*. Chapel Hill, NC: Horizon Research Inc.

Yackel, E., Cobb, P., and Wood, T., 1991. Small-Group interactions as a source of learning opportunities in second-grade mathematics. *Journal for research in mathematics education*, 22 (5), 390–408. doi:10.5951/jresematheduc.22.5.0390.