Exclusively visual analysis of classroom group interactions

Laura Tucker,[1,3,*] Rachel E. Scherr,[2] Todd Zickler,[3] and Eric Mazur[4]

[1]*Department of Physics and Astronomy, University of California,*
Irvine, 4129 Frederick Reines Hall, Irvine, California 92697, USA
[2]*Department of Physics, 3307 Third Avenue West, Seattle Pacific University,*
Seattle, Washington 98119, USA
[3]*School of Engineering and Applied Sciences, Harvard University,*
9 Oxford Street, Cambridge, Massachusetts 02138, USA
[4]*Department of Physics and School of Engineering and Applied Sciences, Harvard University,*
9 Oxford Street, Cambridge, Massachusetts 02138, USA

(Received 5 May 2016; published 30 November 2016)

Large-scale audiovisual data that measure group learning are time consuming to collect and analyze. As an initial step towards scaling qualitative classroom observation, we qualitatively coded classroom video using an established coding scheme with and without its audio cues. We find that interrater reliability is as high when using visual data only—without audio—as when using both visual and audio data to code. Also, interrater reliability is high when comparing use of visual and audio data to visual-only data. We see a small bias to code interactions as group discussion when visual and audio data are used compared with video-only data. This work establishes that meaningful educational observation can be made through visual information alone. Further, it suggests that after initial work to create a coding scheme and validate it in each environment, computer-automated visual coding could drastically increase the breadth of qualitative studies and allow for meaningful educational analysis on a far greater scale.

DOI: 10.1103/PhysRevPhysEducRes.12.020142

I. INTRODUCTION

Classroom observation is an important means of measuring learning behaviors that students and instructors may not accurately self-report in surveys. Analysis of video collected during an activity allows for measure of in-the-moment behavioral response to the classroom environment without bias from retrospection, which can affect recall of emotion [1,2], focus [3], and how time was spent [4,5]. Behavioral data are also important because, unlike high-stakes national and state assessments, both teachers and administrators rate observation as a reliable measure of student learning [6].

Analysis of classroom observations can be broad or detailed in scope, and generally includes conversation along with audible and visible nonverbal cues. Often, classroom observation studies utilize analysis of conversation to measure aspects such as the kinds of explanations given during a group discussion [7,8,9]. Other observational studies use broader measurements such as distribution of time for a given curricular material or the level of student participation [10,11,12]. These broad measurements are important to determine how curricular materials and instructor implementation affect the use of time in student group learning. Both levels of observational study—detailed conversational analysis and broad measurements such as participation level and time use—can provide measurements that are meaningful to educational researchers and practitioners in answering a variety of research questions.

A challenge to both levels of observation research is scalability. With current techniques, it is difficult to balance between two competing needs: the need to do close analysis of a faithful record of the classroom (i.e., video) and the need to reliably code something meaningful quickly. The time barrier of both collecting student-specific audio data and reliably coding audio-visual data restricts the scale of most observation research to small sample sizes, long time requirements for analysis, or expansive coding teams.

Greater scalability would benefit both small and large-classroom observation studies. In small classrooms, scalable techniques for reliable analysis would extend observation capabilities to larger numbers of small-group interactions and over longer time scales. In large classrooms, scalable observation measurement would make it possible to code not only large numbers of students, but dynamic interactions in which student groups change over the course of a single class period. This situation occurs in large-enrollment university courses using techniques such as Peer Instruction. In a class using Peer Instruction,

[*] Corresponding author.
tucker@uci.edu

students are asked to discuss a series of conceptual questions with neighboring students. Students' discussion partners can vary not only between questions, but within a single question a student may take part in discussions with more than one group of students. Enrollment in these classes can reach close to 1000 students. In such conditions, intelligible audio recording of student discussions is logistically difficult not only to collect but also to identify with specific students.

As an initial step towards more scalable classroom group behavior measurement, we use coding protocols, video data, and a previously validated behavioral coding scheme that links students' conversation, nonverbal audio cues, and camera-visible student behaviors to their epistemological framing [13]. In this context, epistemological framing describes students' goals in a learning activity, and the knowledge and reasoning required to achieve those goals [13,14]. For example, epistemology discerns whether a student views the activity as a chance to confront conceptual difficulties of the material, or as an exercise of filling in answers on a worksheet. Epistemology is an important educational aspect because how students view the purpose of an activity mediates their learning [14]. Because of this role, determining students' epistemology for a learning activity is relevant to curriculum designers, instructors, and classroom researchers. While we base our analysis on work that has been validated to code for epistemology, the analysis in this paper is not intended to contribute to research on epistemology specifically. Rather our interest is in coding methods, and our methods could be applied to validated coding schemes linked to other theoretical constructs distinct from epistemology.

Unlike previous uses of this epistemological coding scheme, which used "students' behavior, along with the substance of their speech," we code by using only visual data and compare this to coding with both visual and audio data [13]. Our interest is not in visual markers as an end, but as a proxy for the behavioral codes previously developed. The goal of this research is to present a proof of principle that educationally meaningful assessment of group learning is possible through analysis of visual information alone.

Our coders first code by watching the video without listening to the audio (V mode), then again by watching the video with audio (AV mode). Different coders' results are compared in three ways: V vs V mode, AV vs AV mode, and V vs AV modes. To determine if systematic differences exist between V and AV coding modes, we compare the results of all coders one time interval at a time.

Our results show that interrater reliability is as high when coders watch video without sound as when they watch it with sound. This is true even when comparing coders using both audio and visual data to coders using visual data only; 87% of time intervals qualify as agreement between V and AV modes. In short, visual signals alone are sufficient to reproduce the coding of these behavioral frames.

In the few areas with lower agreement between V and AV coding modes, we find slight differences in the codes assigned in each mode. Visual-only coders (V mode) have a slight inclination towards less interactive codes (e.g., focus on worksheet and listening to the instructor) while AV coders have a slight inclination towards more interactive codes (e.g., group discussion and joking).

With this work as evidence that educationally meaningful qualitative coding can be achieved with visual data, we see the opportunity for significant scaling of qualitative coding through automation using computer vision. Creating coding schemes, as always, requires substantial qualitative work to develop and validate the schemes in the target settings. Then, extending these qualitative studies to a larger scale through computer automation would additionally require testing the reliability of the scheme's visual cues alone. However, after the initial investment, automation would allow the work of many qualitative studies to extend to large-scale assessment at relatively low cost. This has potential to provide valuable feedback to curriculum developers, classroom designers, and instructors across institutions and classrooms.

II. BACKGROUND

The classroom learning activities in our video collection are from *Maryland Tutorials in Physics Sense-Making* [15]. These tutorials are worksheet-based collaborative active-learning activities designed to reconcile students' physics intuition, experimental results, and calculations. Our video captures students working collaboratively in groups of four.

We base our work on the Scherr and Hammer coding scheme, which uses group behavioral clusters that indicate an epistemological frame [13]. These behavioral clusters are identified by speech, other audible cues, and body language that occur simultaneously among group members. Table I breaks down the five behavioral clusters into audio and visual cues [13]. Briefly, in the blue cluster, students focus on completing the tutorial worksheet. In the green cluster, students focus on discussion with their group members. In the red cluster, students focus on interaction with a course instructor or teaching assistant (TA) interacting directly with the group. In the yellow cluster, students engage in emotional or joking behavior. The gray cluster encompasses time in which the students are not involved in small group interaction, including time in which the instructor addresses the entire class. Groups usually transition between different behavioral clusters in unison, allowing behavioral labeling to be done for the group as a whole rather than for each individual [13].

Upon analyzing the types of utterances during each of these behavioral clusters, Scherr and Hammer found that the clusters served to mark not only epistemological framing, but also the conceptual substance of student reasoning. The discussion frame "is associated with original speech about the physics ideas," in contrast to the

TABLE I. Summary of behavioral clusters first identified by Scherr and Hammer [13].

Qualitative code and associated epistemology	Audio cues	Visual cues
Blue: Completing the tutorial worksheet	• Low, quiet, indistinct tone of voice (muttering) • Brief speech episodes • Rising intonation at the beginning and falling intonation at the end of speech episodes • Incomplete sentences of an initiator often completed by a peer	• Eyes primarily on paper • Brief glances at peers • Bodies lean forward 30 degrees to the vertical • Few gestures • Faces relatively neutral
Green: Engaged in a discussion	• Animated voices • Clear, loud, and dynamic speech	• Straight posture (sitting up straight) • Frequent eye contact with one another • Animated faces • Prolific gesturing
Red: Listening to the teaching assistant (TA)	• Little student speech	• Eye contact between students and teaching assistant • Students' bodies straight and still • Few student gestures
Yellow: Joking	• Hedging or joking tone of voice	• Giggling or smiling • Shifting bodies in seats • Touching own face or hair • Unsettled gaze (moves among peers, papers, other points in the room)
Gray: Other		

worksheet frame, which is associated with reading the activity and asking simple, clarifying questions [13]. In further support that these epistemological frames suggest the depth of students' thinking, a study with a subset of this data set showed that 81% of "chaining" (verbally expressed logical inference) occurred in the discussion frame [13].

The Scherr and Hammer coding scheme provides only one perspective of epistemology from a rich literature, much of which focuses on discourse or responses to questionnaires (see, e.g., Refs. [16–18]). We begin with the Scherr and Hammer coding scheme for this analysis because, in addition to analysis of student discourse, it explicitly notes visual cues in addition to speech and other audible cues.

While the original study showed that these behavioral clusters are reliably detectable with a combination of audio and visual data, cues from conversation in particular are rich and defining. Previous work has shown that the conversation and other audio cues together are enough to identify behavioral clusters without visual information [19].

Conversely, as Table I shows, the visual cues that co-occur with the audio cues of each behavioral frame are large in number and well defined. In the blue cluster, students' hands are relatively still (in contrast to gesturing), their faces are neutral, their bodies lean forward towards the table, and their gaze is generally on their papers. There can be eye contact within this label; however, it is fleeting. In the green cluster, there will often be prolific gesturing. Students' faces are animated, they sit up straighter, and they make significant eye contact. In the red cluster, the students also sit up straight like in the green discussion label; however, their gaze direction is heavily favored towards the teaching assistant. The students have little or no gestures. When the students talk, they address their speech to the teaching assistant rather than the group. In the yellow cluster, visual signs of emotion are present such as giggling, smiling, adjusting in seat (straightening or stretching), fidgeting, and touching their face, hair, or arms. They often have an "unsettled gaze:" eyes glancing at multiple places that are distinct from the student's group and his or her paper. Because of the number of visual observations associated with each behavioral cluster, we hypothesized that the coding scheme would have high interrater reliability using only visual information—without audio.

Research into nonverbal communication has explored many of the cues described above including gestures, posture, touching, facial expressions, eye behavior, and nonverbal vocal behavior [20]. A substantial history of work looks at these cues—in isolation from one another—as markers of behavior. Examples include posture congruence indicating involvement [21], gaze behavior marking cognitive processing [22,23], and gestures marking enthusiasm [20]. Combinations of these cues can measure complicated aspects of group dynamics such as group rapport [24] and power within the interaction [25].

Although studies in nonverbal communication often look at interaction in the absence of conversation, the existing notions of "verbal" and "nonverbal" do not perfectly align with our notions of "audio" and "visual." Nonlinguistic audio cues such as speech volume and inflection can be considered nonverbal communication [26]. On the other hand, emblem gestures have a clear linguistic meaning in a culture, such as the placement of a finger in front of the mouth to signal quiet, and can be considered verbal; in fact, such gestures activate the same brain regions as verbal communication [27]. While it is therefore hard to find a perfect match within the sociology literature of an analysis of group dynamics with only visual cues, they do exist. For example, Kendon describes how a specific body movement marks the end of conversation: participants signal the end of a "close salutation" with body movement away from the location of the greeting [28]. Beyond this, some visual nonverbal signals are so strong that computer vision has had success in measuring them with visual cues alone [29].

Though we study video from university physics recitation sections, the general nature of the worksheet activity and visual coding elements suggest that this research is not science specific; we expect visual-only coding should generalize to other coding schemes, and coding schemes beyond epistemology. For example, the Interactive Teaching Map measures on- and off-task classroom behavior and group engagement [11]. Markers such as writing, raising one's hand to speak, and participating in group discussion are similar to markers used in the current study. Similarly, the STROBE tool used to assess medical school education includes a classification of each learner's activity in time, such as "talking, listening, reading, or writing" [12]. Like the current study, measurements like this are likely to have high interrater reliability using only visual cues.

III. METHOD

A. Data set and coding scheme

The data set for this study consists of 886 5-sec intervals (74 min) of video footage; each video captures one group of four students doing a worksheet activity. The complete data set comprises two student groups working on two tutorial worksheet activities.

The data set and coding scheme, using audible and visual cues together, have been used in previous studies and consistently demonstrate high interrater reliability—over 90% renormalized percent agreement (see next section for definition) in the original publication using both visual and aural observations [13]. We describe the original coding scheme in the Sec. II of this paper and show the breakdown of visual and aural cues in Table I.

B. Coding procedure

Similar to the procedure outlined by Scherr and Hammer, coders independently watch each video [13]. For each 5-sec time interval in the video, they apply exactly one of the five codes described in the previous section. However, in our procedure, coders first apply the codes while watching the videos with the sound turned off. We name this mode of coding *visual only*, abbreviated V. The second mode of coding uses both audio and visual information. We name this mode audiovisual, abbreviated AV. The original study used AV coding exclusively.

Four coders participated in our study, each coding the full data set first in V and then in AV mode. The ordering ensures that coders using V mode do not have knowledge of conversation content associated with visual information.

In general, when studies compare two techniques, the design should be symmetric so that the technique first used by a coder is varied. This design alleviates concern that similarity in the results between the techniques is due to coders recalling the data from their initial exposure to the data through coding with the first technique. In this study, however, we do not believe it is possible to code in the V mode after having coded in the AV mode. The AV mode provides coders with speech that coordinates with visual cues, making these visual cues readily associated with the speech even when watching the video again without sound. Coders who have watched the audio with video have a very difficult time, in subsequent watching of the video without sound, unhearing the conversation that they heard previously. In other words, we do not believe it is possible for a coder to first watch the video with sound and then watch it without sound, as the memory of the conversation continues to provide sound to some extent in the coder's mind. On the other hand, the visual cues without sound are not as uniquely memorable. Having seen the visual cues previously does not hamper watching the video subsequently with sound in the same way as using the other order.

C. Coder training

Coders train with an iterative process. First, they read a description of the labeling scheme by Scherr and Hammer [13]. Then they code a short video, which is not part of the data set of this study, from the same collection of video. After completing one training video, new coders compare their results with the results of a trained coder. New and trained coders then discuss areas of discrepancy. If interrater percent agreement is less than 80% on the first training video, new coders repeat the training procedure on additional training video. All coders achieved interrater percent agreement over 80% after one or two training exercises.

D. Analysis

To determine differences in coding between the modes, we ask the following research questions:

(i) How internally reliable is each mode?
(ii) How reliable are the codes when comparing V and AV modes?
(iii) Which code pairs are commonly discrepant between modes? Is there consistent directionality between discrepant pairs of codes?

The most common way to measure interrater reliability is percent agreement. To calculate the percent agreement between two coders, we count the number of time intervals

TABLE II. Examples of code distributions of four coders at single time intervals and their intermode agreement classifications.

	AV mode					V mode				
Intermode agreement	Worksheet	Discussion	TA	Joking	Other	Worksheet	Discussion	TA	joking	other
Perfect code agreement: subset of code agreement where consensus is unanimous	4	0	0	0	0	4	0	0	0	0
Code agreement: both modes come to	3	1	0	0	0	4	0	0	0	0
consensus about a single code	3	1	0	0	0	3	0	1	0	0
Ambiguity agreement: neither mode comes to consensus	2	2	0	0	0	1	2	0	0	1
Code disagreement: both modes	0	4	0	0	0	4	0	0	0	0
come to consensus,	1	3	0	0	0	4	0	0	0	0
but the consensus is different between modes	1	3	0	0	0	3	0	1	0	0
No-consensus disagreement: one mode comes	2	2	0	0	0	3	0	1	0	0
to consensus, but the other does not	0	4	0	0	0	2	1	1	0	0

in a video for which both coders applied the same code. We divide the number of time intervals with consensus by the total number of time intervals to yield the percent agreement between the two coders for that video.

A similar way to measure interrater reliability is renormalized percent agreement, the chief statistic used in the original study [13]. The rationale is that although group behavior overall is strikingly unified, it can take a small period of time for all members of the group to transition between behavioral frames. Renormalized agreement takes this fuzzy transition time into account by ignoring disagreements that last only 5 seconds. A single time interval is removed from the calculation if coders disagree on the code for the single time interval, but agree on the codes for the time interval just before and the time interval just after. With these isolated 5-sec disagreements removed, the calculation of renormalized percent agreement is identical to percent agreement: the total number of consensus lines is divided by the total number of lines.

Another common way to measure interrater reliability is Cohen's kappa [30]. Cohen's kappa is a statistic that accounts for chance agreement in a labeling scheme by taking into account the frequency of the label. Guidelines indicate strength of agreement based on Cohen's kappa values: kappa less than 0 means agreement is less than that expected by chance; kappa above 0.6 is substantial agreement, and kappa over 0.8 is almost-perfect agreement [31].

To determine the interrater reliability of the full labeling scheme, we use percent agreement, renormalized percent agreement, and Cohen's kappa. For each video, we compare different coders two by two in the following mode-comparison categories: (1) V vs V, (2) AV vs AV, and (3) V vs AV. We calculate the mean and standard error of the mean for each category, averaging over all pairs of coding in the category. From these statistics, we measure the interrater reliability within each mode and the reliability between the modes averaged over all raters.

To compare specific similarities and differences in results between V and AV coding modes, we create a code distribution table for each video in each mode. Table II shows an example code distribution table. It shows the number of coders choosing each code at each time interval. It keeps track only of the distribution of codes at each time interval within a single mode, not which coders applied which codes.

We use the following terminology: "consensus" refers to the level of similarity within a single mode, "agreement" refers to the level of similarity between the two modes. Within a mode, consensus occurs when all but one coder apply the same code. Intermode agreement happens in two cases, which Table II illustrates. If both modes have consensus on the same code, the time interval has "code agreement." A special case, "perfect code agreement," occurs if coders in both modes unanimously agree on a single code. If neither mode has consensus, the time interval is considered "ambiguity agreement" because the behavior is ambiguous within each modes. Intermode disagreement happens in two cases. If both modes have consensus but the code is different for each mode, the time interval has "code disagreement." The last type of disagreement arises when there is a consensus in one mode but no consensus in the other mode. This can occur when the consensus is in the V mode only or the AV mode only.

To measure the level of overall agreement, we count the number of lines in each agreement category over both videos. We then analyze the differences in consensus codes in time intervals with disagreement. The next section describes the results of this process.

IV. RESULTS

Using three different metrics, we calculate interrater reliability for both V and AV modes of coding. These results are visible in the first two rows of Table III. Cohen's

TABLE III. Top two rows: Using three different metrics, interrater reliability is high for both V (video only) and AV (audiovisual) modes of coding. Bottom row: Intermode reliability (V vs AV) is high. Bottom two rows: Two coders using different modes (V vs AV) have equivalent interrater reliability as two coders both using AV (same mode). Then, coders' results for both videos are compared two at a time and averaged over coder combinations. Error is the standard error of the mean.

Comparison type (different coders)	Number of time-interval comparisons	Raw agreement (Cohen's kappa)	Raw agreement(%)	Renormalized agreement
V vs V	5316	0.815 ± 0.009	87.4 ± 0.7	91.6 ± 0.7
AV vs AV	5316	0.79 ± 0.01	86 ± 1	89 ± 1
V vs AV	10 632	0.790 ± 0.008	85.7 ± 0.5	89.3 ± 0.6

kappa for V vs V coding has mean value of 0.815, classifying the interrater reliability as near perfect. Reliability of AV vs AV coding is slightly lower with a mean Cohen's kappa value of 0.79, and classifying it as near-perfect agreement within error. The other statistics also show similar levels of agreement within each mode.

The same three metrics applied to intermode comparisons are visible in the bottom row of Table III. In fact, two coders using different modes (V vs AV) have equivalent interrater reliability to different coders both using AV (same mode). Figure 1 also demonstrates the extent of intermode reliability. Combining the results of all coders in all modes, we see that coders between modes agree 87% of the time. Seventy percent of the time, all coders pick the same code such that agreement between modes is perfect. Two percent of the time, both modes agree that behavior is ambiguous. Disagreements in which a consensus code is chosen in each mode but disagrees between modes occur 3% of the time. The final 10% of time involves consensus in one mode only but no consensus in the other mode, with the lack of consensus being in the V and AV modes equally often.

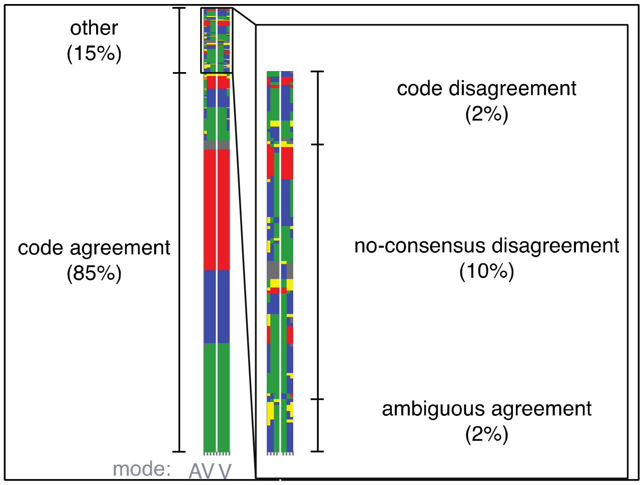

FIG. 1. Intermode reliability. This figure contains the codes, represented by the colors from Table I, for all 886 time segments in the study. Audiovisual (AV) coding is represented in the four columns on the left and visual-only (V) coding in the four columns on the right. Agreement constitutes 87% of time intervals. Explanations of intermode agreement are shown in Table II. The time intervals in this figure are not presented in chronological order; instead, they are reordered and grouped according to agreement type.

The coding modes have small but systematic differences. V coders have a slight inclination towards less interactive codes (e.g., focus on worksheet and listening to the instructor) while AV coders have a slight inclination towards more interactive codes (e.g., group discussion and joking). These differences are evident when looking at code disagreements and no-consensus disagreements in Figs. 2 and 3. As seen in Fig. 2, it is more common for AV coders to code an interaction as discussion when V coders identify the interaction as worksheet than vice versa. Similarly, Fig. 3 shows that there are more time intervals—more than double—for which the AV mode consensus is discussion when the V mode is ambiguous than vice versa. Along the same line, Fig. 3 also shows that there are more time intervals for which the V mode consensus is worksheet when the AV mode is ambiguous than vice versa.

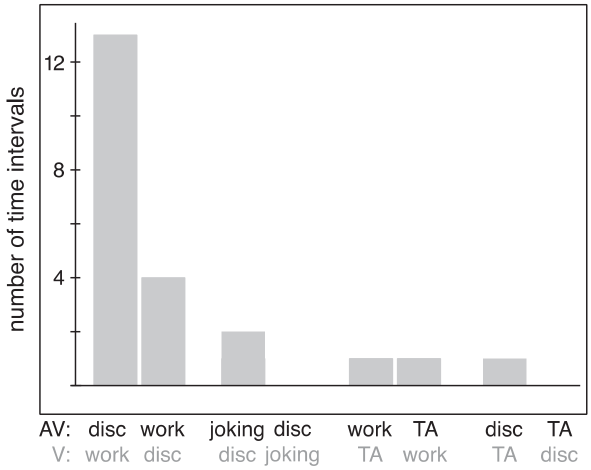

FIG. 2. Code disagreements. Code disagreements occur at time intervals when the codes chosen in the V and AV modes differ. The AV mode coders are more likely to have consensus on discussion when the V mode coders have consensus on worksheet, and vice versa. Also, the AV mode coders are slightly more likely than the V mode coders to recognize an interaction as joking. Combinations of codes not shown above did not have code disagreements.

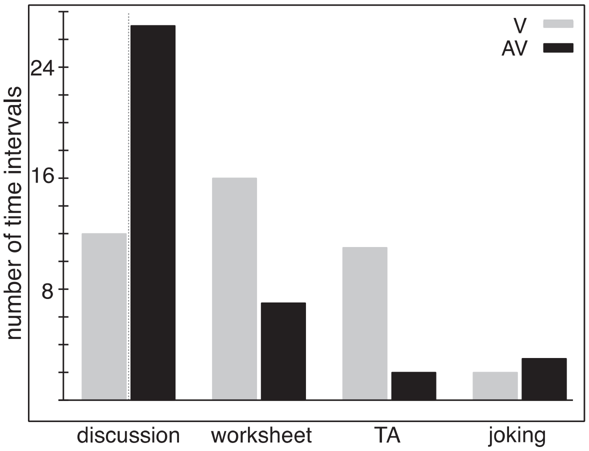

FIG. 3. No-consensus disagreements. No-consensus disagreements occur when one mode has consensus in a time interval when the other mode does not have consensus. The bars show the number of time intervals that the mode showing consensus agreed on a particular code when the other mode did not have consensus. Both the worksheet and discussion bar pairs show that the AV mode is slightly biased toward the discussion code while the V mode is slightly biased toward the worksheet code. Also, the V mode is slightly biased toward the instructor code when behavior is ambiguous in the AV mode.

Figure 3 also shows a stronger bias in the V mode toward the TA code when the AV mode is ambiguous than vice versa. Finally, Fig. 2 also shows that the AV mode has a slight bias towards coding joking over discussion than the V mode.

V. DISCUSSION

The level of reliability both within and between V and AV coding modes is high. This result is even more notable when we consider that, in this example, we reliably reproduce results for an epistemological coding scheme without listening to discourse.

While the overwhelming result is the high reliability between V and AV coding modes, future implementers should be aware of the small differences between the modes. A slight bias exists towards interaction codes (discussion and joking) in the AV mode. This is evident through a combination of results: (1) code disagreements (Fig. 2) between discussion and worksheet show that there are more time intervals where AV coders see behavior as discussion when V coders view behavior as worksheet than vice versa—by over a factor of 3, (2) code disagreements between joking and discussion show that, though the number is small, there are a few time intervals where AV coders see behavior as joking when V coders view behavior as discussion, (3) no-consensus disagreements (Fig. 3) show that AV coders come to consensus on the discussion code when V coders have no consensus twice as often as the reverse. Towards a similar point, a slight bias exists towards lower interaction codes (worksheet and TA) in the V mode. The no-consensus disagreements show a larger number of consensus time intervals in both the worksheet and TA codes in the V mode than in the AV mode.

One possible reason for this bias towards interaction in the AV mode is that in moments of high discrepancy, coders in the AV mode put more reliance on audio cues even when nonverbal cues are present. Previous work suggests that when large discrepancies exist between audio and visual cues, people rely more strongly on audio [32].

It is tempting in this line of research to consider one mode of coding as the correct result against which the other mode is measured. However, neither mode functions as "ground truth." AV coders have access to a rich and large set of information from both audio and visual data. It is possible a complete data set including both audio and visual data yields the most accurate behavioral information. It is also possible that some of this information distracts from visual behavioral indicators so that the audio cues are weighted too heavily in the decision to apply a given code.

A related illustration comes from musical competitions. Tsay studied judgment of musical competitions by musically trained and musically novice study participants [33]. Although the vast majority of participants, when asked, considered sound to be the most important feature of a musical performance, participants who watched the performances without audio were far more likely to select the winner than participants who listened without visual data or both watched and listened to the video. Although participants considered sound as the most important determination in choosing a musical winner, visual cues seemed to be the dominant factor in judgment. Similarly, although we might think of conversations as defining group interaction, visual behaviors may give an equal or more reliable read on educationally significant behaviors.

Increasingly, the extensive manual effort required to train and deploy V coders can be avoided by instead using computer vision systems that automatically recognize activities and interactions from video. These computer vision systems are "trained" by being given a set of exemplar video clips for each activity or interaction category, and then they are deployed to recognize and annotate instances of those categories in new, unlabeled video. Ongoing advances in computer vision are allowing reliable performance in a growing diversity of environments, using a growing variety of visual cues such as gaze directions, body positions, and specific movements [34–37]. In fact, recent work suggests that large classrooms are particularly amenable to such automated V coding, since students remain relatively localized around their seating positions [38].

In regards to the coding process, coders reported that AV coding is easier than V coding. We speculate the reason for this is that in AV coding, some behavior is duplicated by vocal and visual cues, making it easier to notice. For example, some information is highlighted by small vocal responses that an observer using both audio and video would catch effortlessly. This is likely because, in contrast to an activity like a basketball game in which the ball is an obvious indicator of action, action in a conversation is less obvious visually. Without audio, it is easy for an observer to focus on one area of the screen where the previous action occurred and therefore miss what is happening in other areas. However, we note that while continually surveying a group as a whole is challenging for a human, it would not be difficult for a machine.

VI. CONCLUSION

Visual coding has high interrater reliability both between visual-only coders and when comparing visual-only coders with coders using both audio and visual data. Because of this result, we encourage implementation of visual coding for use with established coding schemes that have been tested for reliability between visual and audio-visual coding. In particular, we see this result as advancing the amount of observation possible in environments that make it difficult to record student-specific audio. We also see this result as expanding the number of student groups that can be analyzed in small classrooms and making studies of interaction in large-enrollment classrooms more tractable.

When establishing reliability between visual-only and audio-visual coding modes of coding in a scheme, it is important to be aware of differences between audio-visual and visual-only coding. Audio-visual coders are likely to rely more on audio cues than on visual cues in times when behavior is highly ambiguous.

Though the work in this paper uses human coders, advances in computer vision open the door to fully automated analysis of student classroom video. This would allow for empirically driven measures of classroom behavior on a large scale. Such measures would be of tremendous utility to analyze with statistical power collaborative classroom activities in a given curriculum, pedagogy, or physical environment.

ACKNOWLEDGMENTS

Several people contributed to the work described in this paper. L. T. and R. S. conceived of the basic idea for this work. L. T. designed and carried out the experiments, and analyzed the results. R. S., T. Z., and E. M. supervised the research and the development of the manuscript. L. T. wrote the first draft of the manuscript; all authors subsequently took part in the revision process and approved the final copy of the manuscript. Orad Reshef and Kelly Miller provided feedback on the manuscript throughout its development. Abdi Aden, Marinna Madrid, Ryan Smith, and L. T. coded video. The research described in this paper was supported by The National Science Foundation under Contract No. NSF IIS-0835338 and The Harvard Initiative for Learning and Teaching.

[1] L. F. Barrett, The relationships among momentary emotion experiences, personality descriptions, and retrospective ratings of emotion, Personality Soc. Psych. Bull. **23**, 1100 (1997).

[2] C. N. Scollon, E. Diener, S. Oishi, and R. Biswas-Diener, Emotions across cultures and methods, Journal of cross-cultural psychology **35**, 304 (2004).

[3] E. Babad in *The New Handbook of Methods in Nonverbal Behavior Research*, edited by J. A. Harrigan, R. Rosenthal, and K. R. Scherer (Oxford University Press, New York, 2008), p. 283.

[4] T. C. M. Lam and P. Bengo, A comparison of three retrospective self-reporting methods of measuring change in instructional practice, Am. J. Eval. **24**, 65 (2003).

[5] D. P. Mayer, Measuring instructional practice: can policymakers trust survey data?, Educ. Eval. Policy Anal. **21**, 29 (1999).

[6] T. R. Guskey, Multiple sources of evidence: an analysis of stakeholders' perceptions of various indicators of student learning, Educ. Meas. **26**, 19 (2007).

[7] M. C. James and S. Willoughby, Listening to student conversations during clicker questions: What you have not heard might surprise you!, Am. J. Phys. **79**, 123 (2011).

[8] N. M. Webb, M. L. Franke, T. De, A. G. Chan, D. Freund, P. Shein, and D. K. Melkonian, 'Explain to your partner': teachers' instructional practices and students' dialogue in small groups, Cambridge J. Educ. **39**, 49 (2009).

[9] E. Woodruff and K. Meyer, Explanations from intra- and inter-group discourse: Students building knowledge in the science classroom, Res. Sci. Educ. **27**, 25 (1997).

[10] I. S. Carvalho and B. Williams, Facilitating the adoption and dissemination of active learning methodologies in engineering education using a classroom observation approach, in *SEFI 37th Annual Conference* (Rotterdam, Netherlands, 2009), p. 1.

[11] D. Kerr, L. Kent, and T. C. M. Lam, Measuring program implementation with a classroom observation instrument, Eval. Rev. **9**, 461 (1985).

[12] P. A. Kelly, P. Haidet, V. Schneider, N. Searle, C. L. Seidel, and B. F. Richards, A comparison of in-class learner engagement across lecture, problem-based learning, and team learning using the STROBE classroom observation tool, Teach. Learning in Med. **17**, 112 (2005).

[13] R. E. Scherr and D. Hammer, sStudent behavior and epistemological framing: examples from collaborative active-learning activities in physics, Cognit. Instr. **27**, 147 (2009).

[14] A. Elby, Helping physics students learn how to learn, Am. J. Phys. **69**, S54 (2001).

[15] A. Elby, R. Scherr, T. McCaskey, R. Hodges, E. F. Redish, D. Hammer, and T. Bing, *Maryland Tutorials in Physics Sense-Making* (Open Source DVD, 2007).

[16] A. A. diSessa, Toward an epistemology of physics, Cognit. Instr. **10**, 105 (1993).

[17] D. Hammer and A. Elby, in *Personal Epistemology: The Psychology of Beliefs about Knowledge and Knowing*, edited by B. K. Hofer and P. R. Pintrich (Routledge, New York, 2002).

[18] B. K. Hofer, Dimensionality and disciplinary differences in personal epistemology, Contemp. Educ. Psychol. **25**, 378 (2000).

[19] M. A. Martinuk, Ph.D. thesis, The University of British Columbia, 2012, http://hdl.handle.net/2429/43082.

[20] M. L. Knapp and J. A. Hall, *Nonverbal Communication in Human Interaction*, 7th ed. (Cengage Learning, Boston, 2010).

[21] M. Lafrance and M. Broadbent, Group rapport: Posture sharing as a nonverbal indicator, Group Org. Management **1**, 328 (1976).

[22] G. Doherty-Sneddon, F. G. Phelps, and L. Calderwood, Gaze aversion during children's transient knowledge and learning, Cognit. Instr. **27**, 225 (2009).

[23] A. McCarthy, K. Lee, S. Itakura, and D. W. Muir, Cultural display rules drive eye gaze during thinking, Journal of cross-cultural psychology **37**, 717 (2006).

[24] L. Tickle-Degnen and R. Rosenthal, Group rapport and nonverbal behavior, Group Processes & Intergroup Relations **9**, 113 (1987).

[25] N. E. Dunbar and J. K. Burgoon, Perceptions of power and interactional dominance in interpersonal relationships, J. Social Personal Relationships **22**, 207 (2005).

[26] S. Patel and K. Scherer, in *Communication*, edited by J. A. Hall and M. L. Knapp (Hubert & Co., Berlin, 2013), Vol. 2, p. 167.

[27] J. Xu, P. J. Gannon, K. Emmorey, J. F. Smith, and A. R. Braun, Symbolic gestures and spoken language are processed by a common neural system, Proc. Natl. Acad. Sci. U.S.A. **106**, 20664 (2009).

[28] A. Kendon, *Conducting Interaction: Patterns of Behavior in Focused Encounters* (Cambridge University Press, New York, 1990).

[29] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, Modeling dominance in group conversations using nonverbal activity cues, IEEE Trans. Audio, Speech, Lang. Proc. **17**, 501 (2009).

[30] K. A. Hallgren, Computing inter-rater reliability for observational data: an overview and tutorial, Tutorials in Quant. Methods Psych. **8**, 23 (2012).

[31] J. R. Landis and G. G. Koch, A Biometrics Invited Paper. Estimation of "Safe Doses" in Carcinogenic Experiments, Biometrics **33**, 1 (1977).

[32] B. M. DePaulo, R. Rosenthal, R. Eisenstat, P. L. Rogers, and S. Finkelstein, Decoding discrepant nonverbal cues, J. Personality Soc. Psych. **36**, 313 (1978).

[33] C. J. Tsay, Sight over sound in the judgment of music performance, Proc. Natl. Acad. Sci. U.S.A. **110**, 14580 (2013).

[34] J. K. Aggarwal and M. S. Ryoo, Human activity analysis: A review, ACM Comput. Surv. **43**, 1 (2011).

[35] W. Choi and S. Savarese, Understanding collective activities of people from videos, IEEE Trans. Pattern Anal. Mach. Intell. **36**, 1242 (2014).

[36] A. Hakeem and M. Shah, Learning, detection and representation of multi-agent events in videos, Artif. Intell. **171**, 586 (2007).

[37] T. Lan, Y. Wang, W. Yang, S. N. Robinovitchand, and G. Mori, Discriminative latent models for recognizing contextual group activities, IEEE Trans. Pattern Anal. Mach. Intell. **34**, 1549 (2012).

[38] R. Li, P. Porfilio, and T. Zickler, Finding group interactions in social clutter, in *IEEE Conference on Computer Vision and Pattern Recognition* (Oregon Convention Center, Portland, 2013), p. 2722.