Article

Organizational Research Methods
1-16
© The Author(s) 2016
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1094428116647785
orm.sagepub.com
SAGE

Systematic Behavioral Observation for Emergent Team Phenomena: Key Considerations for Quantitative Video-Based Approaches

Mary J. Waller[1] and Seth A. Kaplan[2]

Abstract

The use of moving images to generate data for behavioral analysis has long been a methodology available to organizational researchers. In this article, we draw from previous research in team dynamics to describe and discuss various methodological approaches to using video recorded behavior as a source of quantitative data. More specifically, we identify and examine key decision points for researchers and illustrate benefits and drawbacks to consider. The article concludes with suggestions for ways in which quantitative video-based approaches could be improved.

Keywords

behavioral observation, video data, audio data, coding techniques, team dynamics, behavioral simulation, team training

In the early 20th century, the advent of Frederick W. Taylor's conceptualization of scientific management introduced the notion of time studies to the, at that time, new and growing field of organizational management. Time studies used direct observation and analysis of workers' movements over time as they performed tasks, the purpose being to derive the most productive "one best way" of performing the repetitive physical movements necessary to complete the tasks (Nelson, 1980). Building on Taylor's time studies, Frank and Lillian Gilbreth introduced the use of film in motion studies that visually recorded workers' movements as well as recording the timing of movement occurrence (Lancaster, 2004). While the work of both Taylor and the Gilbreths was roundly criticized on various grounds ranging from subjectivity to dehumanization, their methods helped create a solid foundation for observational studies of work behavior in organizations.

[1] Schulich School of Business, York University, Toronto, Ontario, Canada
[2] Department of Psychology, George Mason University, Fairfax, VA, USA

Corresponding Author:
Mary J. Waller, Schulich School of Business, York University, Toronto, Ontario, Canada.
Email: mwaller@yorku.ca

Today, much of that work behavior has shifted from skilled individuals engaging in discrete repetitive movements to knowledge work requiring the gathering, sharing, and interpretation of complex information by *teams* of individuals, often acting under time pressure and on behalf of their organizations (Kozlowski & Bell, 2003). Rather than observing minute hand movements and posture as individuals manipulate tools and machinery, many social science researchers now focus intently on the interactions of team members as they struggle to make sense of unfolding situations and choose appropriate actions to take. The nuances and subtleties of these real-time behaviors and interactions can be captured for later analysis by a variety of digital video recording devices that seem to continually increase in technological quality and decrease in price, making the exploration of emergent phenomena and interaction patterns in teams—aspects of interaction typically unnoticed by the casual observer—quite possible. As George C. Homans (1951) remarked in his seminal book *The Human Group*: "There is still only one sufficient reason for studying the group: the sheer beauty of the subject and the delight in bringing out the formal relationships that lie within the apparent confusion of everyday behavior" (p. 454). Although video-based data and subsequent statistical analyses help researchers work toward uncovering these relationships, the overarching research goal—identifying and understanding the influence of team-level phenomena—remains remarkably similar to the goal of Homans and his contemporaries. Indeed, many of us still believe that groups and teams are complex and confusing, yet beautiful, entities.

In this article, we describe using video recordings to generate quantitative data for the study of team dynamics—that is, the study of behavioral phenomena that emerge as teams of interdependent members work over time toward a common goal or outcome. Drawing from work in this area, we highlight key decision areas across various approaches. Throughout this exploration, we use information concerning one of our own studies to illustrate decisions made and identify specific challenges for researchers to consider. The decision areas we discuss here concern field data collection, coding schemes and intervals, coder selection and training, and analyses. It should be noted that these decision areas, depicted in Figure 1, are interdependent; for example, the nature of the data collection site may determine the quality of the video; this in turn may affect which behaviors can be reliably coded into usable data. Similarly, the need to capture a certain frequency of the behaviors of interest may determine the choice of data collection site, the coding approach, and ultimately, the analysis used.

Decision Area 1: Data Collection Site

The use of video recordings as a source of research data necessitates that the right to consent has not been violated and that the right to confidentiality is rigorously maintained not only during video capture but also throughout the entire research process (Israel, 2014). Institutional human subjects committees may require specific steps be taken to acquire individuals' consent, store video recorded data, de-identify (i.e., labeling data not by participants' identities but rather by unique reference numbers or codes) any coded data derived from video recordings, and destroy video recordings by a certain date. In many organizational contexts, union approval must be obtained as well; due to concerns about the privacy of their members, union consent may be comparatively more difficult to obtain when video data, versus survey-based or other types of data, are involved. Assuming that institutional, organizational, and union guidelines can be met, video recording devices can capture specific phenomena of interest that are unpredictable in either onset, duration, or both—unpredictability that makes timing direct researcher observation difficult or impossible.

In addition to relieving researchers from the requirement of constant direct observation, video devices are continually becoming less obtrusive and expensive, allowing for the video capture of work behavior in places physically difficult for a researcher to observe work behaviors. The Polaroid Cube, for example, measures 35 mm on each side, includes a rubbery exterior and a rechargeable battery, records 124° wide-angle high definition video on a Micro SD card, has a magnetic base for

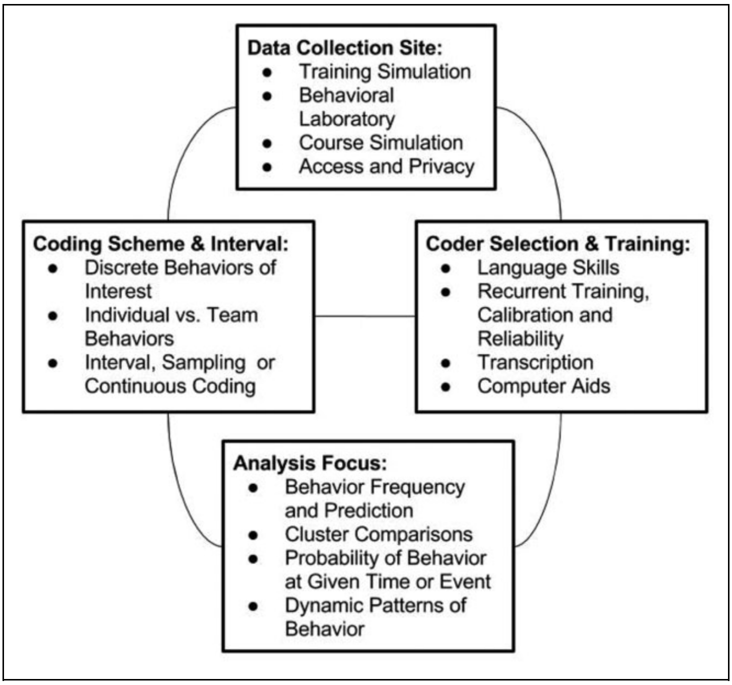
Data Collection Site:
• Training Simulation
• Behavioral Laboratory
• Course Simulation
• Access and Privacy
Coding Scheme & Interval:
• Discrete Behaviors of Interest
• Individual vs. Team Behaviors
• Interval, Sampling or Continuous Coding
Coder Selection & Training:
• Language Skills
• Recurrent Training, Calibration and Reliability
• Transcription
• Computer Aids
Analysis Focus:
• Behavior Frequency and Prediction
• Cluster Comparisons
• Probability of Behavior at Given Time or Event
• Dynamic Patterns of Behavior

Figure 1. Key aspects of video-based behavioral observation studies.

easy attachment, and currently costs under $100 USD. The reduced size and price of such devices allows multiple cameras to be easily placed in work spaces, including those in difficult environments such as hospitals (Su et al., 2015) or heavy equipment cabs. Additionally, many organizations use video cameras in their daily operations for monitoring, safety, and security reasons; again assuming that proper consent can be obtained, existing video systems ranging from wearable devices to dash cams could be used to gather video recordings of behavior.

An alternative method for creating video recordings of work behaviors involves organizations' existing training and assessment efforts. Organizational training departments can be a source of video recorded work behavior, albeit in a simulated work environment. Video (or audio if video is not available) recordings are often created during the recurrent simulation training that many teams participate in as required by organizations, industries, or regulatory agencies. Far from a game-like setting, these simulations often take place in extremely expensive, realistic settings such as full-motion flight simulators, plant control room simulators, or crisis command centers and are taken very seriously by team members; both individual and team performance during the simulations are often used by organizations for evaluative purposes (Stanton, 1996). Many such training facilities are equipped with multiple video cameras and microphones and use video and/or audio recordings during team debriefing and assessment activities; however, organizations typically do not use these recordings to produce data for in-depth quantitative behavioral analyses. By using recordings that an organization is already producing, little extra time or effort on the part of training staff is required, often making the prospect of granting access to researchers more palatable for a participating organization.

Our Example

Whether recording work behaviors in real or simulated environments, initial meetings with organization representatives are critical in terms of developing research questions that both relate to existing theory and pertain to practical issues. We typically begin a study by developing a few overarching research questions derived from existing literature and observations and then speak at length with team members as well as the training professionals who design and implement the simulation scenarios the teams will face. During these conversations, we refine our research questions and focus and try very hard to identify questions that are compelling to both the training staff and a potential academic audience, with the understanding that our research results will be shared directly with the organization.

As our example here, we refer to our study of nuclear power plant control room crews (Stachowski, Kaplan, & Waller, 2009) in which we examined how characteristics of control room crews' (i.e., teams') communication patterns related to crew overall performance in high fidelity simulations. The primary research question of interest was whether more patterned (or consistent) versus less patterned team communication was associated with superior team performance. Prior research supported both perspectives, and as such, we did not offer hypotheses in the study but rather approached the question in a more exploratory manner.

The study was conducted at a nuclear power plant site in the United States. Worth noting is that the utility company we partnered with for this study was quite supportive of the research, making for a very effective collaboration. We were initially contacted by the training manager who had learned of our previous research in this area; we subsequently worked with this manager—our "champion" within the organization—to identify questions to be investigated that were specific and meaningful for the plant. The manager was also instrumental in introducing us to key decision makers in the plant. We were invited to present our research design to the top management team of the corporation that owned the plant; following this team's approval of the research plan, we worked closely with training supervisors to schedule data collection. We relied on simulation/training coordinators to furnish technical information (and often documentation) about their video recorded simulations—such as explaining to us the nature of the nonroutine events the teams were addressing and deciphering the technical language that team members used. Having this support from the study participants was important given that per our human subjects committees' guidelines, we needed all members of a team to consent to participate in the study; if even one team member had chosen not to consent to our using his or her team's video recording for research, we could not use the video of that entire team's simulation training.

In this example study, we used as a source of data the video recordings of 14 intact nuclear power control room crews as they responded to a series of nonroutine events during regularly scheduled simulation training. Each crew was composed of a supervisor (the team leader), at least two systems operators, and typically at least one additional member. In this context (but not all that we have studied), the team members had very specific and clearly defined roles and responsibilities. The simulations took place in a control room that was an exact replica of the one in which a crew normally worked. All simulations at the plant were video recorded and then typically used for debriefing and training purposes. Four digital video cameras (which also recorded audio) were already installed throughout each simulation control room and recorded simultaneously. Having the multiple cameras was necessary to capture all of the crew members as they moved throughout the control room during the simulation. Having these multiple cameras also provided more than one perspective (i.e., images from more than one camera) of the same crew member. This redundancy proved very beneficial during the coding process as the coders were better able to determine who was talking to (and responding to) whom.

During the simulations, the crews responded to multiple nonroutine, crisis-like events that were presented sequentially. The training personnel at the plant had written the scenarios, some of which reflected actual events that recently had occurred at other plants. Each simulation lasted at least an hour, with some lasting up to almost three hours. We decided to focus on team behavior during the first simulated crisis event. Doing so avoided the potential for earlier performance to impact performance on subsequent events. Furthermore, based on our observations of the simulations and our review of relevant literature, we had learned that much of the team interaction and the key decision making occurs within the initial 15 minutes after the onset of the event, so we focused our behavioral observation efforts on this period.

Decision Area 2: Coding Schemes and Intervals

Coding schemes used to categorize and record occurrences of specific discrete behaviors began appearing in the group dynamics literature long ago; one of the earliest and most widely used methods is Bales's (1950) coding scheme for the interaction process analysis (IPA) approach. A coding scheme is a set of rules used to assign a category label or "code" to an observed instance of a target behavior. In direct observation, observers may watch individuals or teams engaging in behaviors and code target behaviors—that is, note on a paper or electronic table the time at which the behavior occurred and the predefined category within which the behavior fell. This same coding process may also be used with video recorded behaviors. *Microcoding* (see e.g., Stoolmiller, Eddy, & Reid, 2000) is also a term used for noting the timing and frequency of certain behaviors captured on video or audio recordings; others may use different terminology for the same process.

In general, coding schemes of group and team behavior can vary along various dimensions, such as the number of behaviors, the degree to which the scheme is generic versus tied to a specific context or task, and the level of abstraction of the coding (for a more general discussion and set of recommendations regarding these dimensions for coding group processes, see Weingart, 1997). To illustrate these differences, we contrast the schemes from two studies of team crisis response in medicine. An example of a more specific coding scheme with fewer behaviors comes from a study done by Marsch and colleagues (2005). The main purpose of this study was to examine adherence to algorithms of cardiopulmonary resuscitation (CPR) among first responders during simulated cardiac arrests. Given this specific objective, these authors decided to code a small set of behaviors particular to that setting (first diagnosis of cardiac arrest, calling of a code, first defibrillation, start of cardiac massage). One can contrast this approach with the use of a more inclusive scheme such as the Co-ACT approach, developed by Kolbe, Burtscher, and Manser (2013). The authors developed this scheme to be more general (focusing on team coordination) and chose 12 categories of behavior based on a theoretical framework of coordination. As such, this scheme can be applied to different contexts within acute care (e.g., Kolbe et al., 2014).

As these contrasting examples make clear, schemes can vary considerably. Given the diversity of coding schemes, how then should researchers develop or choose one? First, unless one is interested in coding behaviors unique to that particular context (like starting cardiac massage in the Marsh and colleagues, 2005, study), researchers might consider using an existing scheme as a "starting point." Table 1 depicts a selection of references for existing schemes. Often, the same sets of behaviors are common and consequential across team contexts and scenarios (e.g., providing information, giving commands, requesting information). After deciding on a scheme, one then can tailor it to the specific context by adding some behaviors or ignoring others. As an example of this approach, Fernandez Castelao and co-authors (2011) borrowed from Kolbe and colleagues' (2013) scheme and then supplemented it with additional behaviors (e.g., those relevant to leadership and to the specific CPR context).

Beyond choosing which behaviors to include, the other major decision regarding coding is that of determining at how granular a level to code behaviors and to what more general category to assign behaviors. Thus, for instance, if a nurse informs a physician "I started CPR," one alternatively could code this as "providing information" or as "nurse providing physician with information about a beginning procedure on a patient." The former approach may be sufficient, but different questions or

Table 1. Resources Coding Schemes, Coder Training, and Statistical Analyses for Video Recordings of Teams.

Developing coding schemes
Kauffeld, S., & Lehmann-Willenbrock, N. (2012). Meetings matter: Effects of team meetings on team and organizational success. *Small Group Research*, *43*, 130-158.
Kolbe, M., Burtscher, M., & Manser, T. (2013). Co-ACT—A framework for observing coordination behavior in acute care teams. *BMJ Quality & Safety*, *22*, 596-605.
Kolbe, M., Strack, M., Stein, A., & Boos, M. (2011). Observing coordination in human group decision-making: MICRO-CO—A micro-analytical taxonomy for analysis of coordination mechanism in decision-making groups. In M. Boos, M. Kolbe, P. Kappeler, & T. Ellwart (Eds.), *Coordination in human and primate groups* (pp. 199-219). Heidelberg: Springer.
Weick, K. (1985). Systematic observation methods. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology. Vol. 1, Theory and method* (pp. 567-634). New York, NY: Random House.
Weingart L. R. (1997). How did they do that? The ways and means of studying group process. *Research in Organizational Behavior*, *19*, 189-239.
Coder training and calibration
Lehmann-Willenbrock, N., Meinecke, A. L., Rowold, J., & Kauffeld, S. (2015). How transformational leadership works during team interactions: A behavioral process analysis. *Leadership Quarterly*, *26*, 1017-1033.
Stachowski, A., Kaplan, S. A., & Waller, M. J. (2009). The benefits of flexible team interaction during crises. *Journal of Applied Psychology*, *94*, 1536-1543.
Statistical analyses
<u>Multiple regression</u>
Westli, H. K., Johnsen, B. H., Eid, J., Rasten, I., & Brattebo, G. (2010). Teamwork skills, shared mental models, and performance in simulated trauma teams: An independent group design. *Scandinavian Journal of Trauma Resuscitation, and Emergency Medicine*, *18*, 47.
<u>Conditional likelihood logit models</u>
Waller, M. J. (1999). The timing of adaptive group responses to non-routine events. *Academy of Management Journal*, *42*, 127-137.
<u>Lagged sequential analyses</u>
Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. New York, NY: Cambridge University Press.
Kolbe et al. (2014). Monitoring and talking to the room: Autochthonous coordination patterns in team interaction and performance. *Journal of Applied Psychology*, *99*, 1254-1267.
<u>Neural network approach</u>
Kennedy, D. M., & McComb, S. A. (2014). When teams shift among processes: Insights from simulation and optimization. *Journal of Applied Psychology*, *99*, 784-815.
<u>Theme software pattern recognition algorithm</u>
Lei, Z., Waller, M. J., Hagan, J., & Kaplan, S. Team adaptiveness in dynamic contexts: Contextualizing the roles of interaction patterns and in-process planning. *Group and Organization Management*. Advance online publication. doi:10.1177/1059601115615246

foci may require the latter, more specific coding. Obviously, with the second approach, one still then can combine these specific behaviors to a more general "provides information" category.

In addition to choosing which behaviors to code, there are also decisions to make regarding coding *intervals*. If target behaviors are easily identified directly from video recordings, interval coding may be appropriate—that is, identifying an interval of time within which human coders watching the video are able to note the occurrence of any of the target behaviors. For example, several studies of dyadic or team interaction use coding intervals of 10 seconds (e.g., Lim & Murnighan, 1994; Waller, 1999). Coders watch or listen to recordings in predefined 10-second increments, noting either the binary occurrence or frequency of target behaviors. Although this technique increases the granularity of the coding, it also increases the ease and speed at which coders complete and compare their work. Weick (1985) suggests to set the interval length so it is long enough to capture one complete instance of any of our target behaviors but not so long as to likely capture more than one instance.

However, if target behaviors on audio or video recordings are difficult for coders to hear or see, professional transcription services may be used to transcribe the recordings and note both time and speaker for each utterance, statement, or thought unit. Coders then use continuous coding (see Kolbe & Boos, 2009) rather than interval coding, assigning a code to each utterance (or other type of word grouping) on each transcript. This event coding does not aggregate utterances into intervals but instead assigns a code to each "event" or utterance on the transcript in a continuous manner. Overall, transcription helps facilitate coding verbal behaviors and some characteristics of interaction (e.g., turn-taking or interruptions) but may be of limited or no use for coding nonverbal behaviors such as body language.

A third possibility in terms of choosing coding intervals involves coding by sampling for specific behaviors, either at scheduled or random points, throughout a video or audio recording. This approach might be useful in the case of studies designed to track or describe patterns of behavior over long periods of time (see Noldus, Trienes, Hendriksen, Jansen, & Jansen, 2000). Sampling can also be used in conjunction with the occurrence of covariants; for example, sampling a certain time interval before and after the occurrence of a nonroutine event might provide information concerning target behaviors associated with such events.

Our Example

Before commencing our study of nuclear power plant control crews, one of us spent several days at the site becoming familiar with the actual team context. We have learned that observing simulations and speaking with key personnel before beginning data collection is essential, especially for the purposes of revising research questions to suit that context and for choosing the behaviors to code. Here, these initial observations and discussions greatly enhanced our understanding of the tasks the teams faced and in turn the behaviors that were most important and frequent in carrying out those tasks—and that we therefore should code. We concluded that the primary tasks for these crews involved diagnosing the underlying cause(s) of the abnormal system indicators (and corresponding alarms) and based on the (evolving) diagnosis, implementing standardized procedures to redress the root causes. Achieving these tasks involved the crew members sharing information with the crew supervisor who then had to incorporate this information, choose which restorative procedures to implement, and coordinate the crew as it enacted those procedures. Essentially, the crews needed to gather and share information and solve problems.

Based on this determination of primary tasks as well as observation of three actual videos, we chose and defined an initial set of behaviors that seemed critical. We then discussed this tentative coding scheme with the training coordinator from the plant and made modifications based on his feedback. The final coding scheme consisted of 11 behaviors. Table 2 displays the master coding protocol that we developed and then used when coding the videos.

As emphasized previously, researchers need to make several key decisions in developing and using schemes to code group behavior (see Weingart, 1997). We discuss three instances that emerged in our study to illustrate this point. First, we learned that during training simulations, the crews respond to multiple nonroutine (i.e., crisis-like events) presented sequentially. For reasons

Table 2. Master Coding Protocol from Stachowski, Kaplan, and Waller (2009) Article on Nuclear Power Plant Control Room Crews.

Coded Behavior	Definition	Examples	Related Variables
Provides information	Supervisor or operator provides unsolicited information that does not follow a request	"I am closing the valve." "Larry, if X or Y occurs, we will trip the reactor."	Information Processing Situation Awareness Threat Recognition
Provides summary/ recap	Supervisor summarizes what has, is, and/or may transpire	"OK, we've completed the first 7 steps."	Situation Awareness SMM Development
Provides feedback	Supervisor or operator gives other team member(s) positive or negative feedback.	"I'm sorry." . . . "That's ok," "Nice job."	Psychological safety
Makes command	Supervisors provides task or instruction to operator(s)	"Larry & Matt, refer to 156 for downfire." "Larry, commence forcing sprays."	Information Processing Resource Allocation
Offers opinion *does not include just general assessments	Supervisor or operator announces his or her opinion regarding the task/situation	"I'm not sure I agree; I think we need to . . . "	SMM Development Situation Awareness
Begins procedure	Supervisor initiates a standard procedure in order to address the nonroutine event Coded when the supervisor begins progressing through the steps of the procedure.	Supervisor will say something like, "We're in #1243."	
Expresses warning	Supervisor or operator offers information that does not follow a request and that clearly indicates something is wrong	"I think we have worsening vacuum." "RCS pressure is 1,900 lbs and lowering."	Threat Recognition Situation Awareness
Pacing	Supervisor comments on or adjusts the current or future pacing of team members	"Ok, let's take our time here."	Prioritization
Requests opinion	Supervisor or operator asks opinion of other(s)	"What do you think?" "Fix this?"	Problem Solving Participative DM
Shift manager returns	Shift manager returns to team after communicating with outside stakeholder	[Coded from behavior only, not utterances]	Interruption Resource Allocation
Begins/ends focus brief	Supervisor holds a briefing with the team	"So, Matt and Larry, focus brief, 2575, before we X, we will Y . . . "	Mental model adaptation

explained previously, we decided to focus on team behavior during the first simulated crisis event and particularly on the initial 15 minutes after the onset of the event.

Second, we learned that after the crew supervisor had initiated a procedure to address a given nonroutine event, much of the subsequent communication consisted of the crewmembers progressing through a series of standardized tasks and checklists. Because the content of this communication was prescribed by the procedure, we did not code it, instead only coding what we determined as "volitional" utterances—those that were not prescribed by the procedure.

Finally, we also learned that almost without fail, the crews used closed loop communication. So, for example, an operator (José) might state, "Ron, I have completed step 2"; the supervisor (Ron) would affirm, "José, you completed step 2"; and Jose then would close the loop with, "That's correct." Because the operators adhered to the use of closed loop communication so strictly, certain types of communication errors—such as one team member failing to hear critical information that was expressed to him—were rare, and we did not code for them. In other contexts we have studied, team members regularly will "talk to the room" (see Kolbe et al., 2014) instead of to a particular individual and will fail to acknowledge appropriate receipt of that message. In those cases, information gets "lost" or distorted much more frequently, and oftentimes, these lapses result in significant adverse events. Thus, in those contexts, we not only would record communication errors, but we also likely would code the rare use of more constructive types of communication (e.g., when the speaker does identify the intended recipient of the message). In sum, one must become very familiar with the communication norms of that particular context and of the specific tasks and types of utterances that occur and matter for each context.

Decision Area 3: Coder Selection and Training

Until a reasonable alternative appears in terms of applicability, affordability, and reliability, most researchers using the behavioral observation approaches described here will continue to use human coders to generate the quantitative data from the audio and video recordings used for analyses. Having individuals who are native speakers of the language spoken on the recordings is often important in terms of understanding colloquialisms and is especially useful for coding recordings from settings presenting very loud ambient noise or alarms that make understanding what is spoken (or yelled) by team members difficult.

As the coders are essentially the researcher's measurement "instruments" via coding, depending on the focus of the research, those coders who have studied psychology or organizational behavior may bring a certain level of precision to the task when it comes to matching category codes to team members' verbalizations or other behaviors. Some studies have specifically used subject matter experts to do the coding. For example, DeVita, Schaefer, Lutz, Wang, and Dongilli (2005) had professional trainers perform the coding. Alternatively, Xiao, Hunter, Mackenzie, Jefferies, and Horst (1996) asked job incumbents who were expert with respect to the task (resuscitations) to perform the coding. We also have found that having access to these experts is beneficial, at least to consult when coding questions arise (e.g., regarding the use of technical jargon). Table 1 provides references for conducting thorough coder training.

An additional aid for coders is software that helps automate the coding process to a certain extent. For example, we have used Noldus's Observer (www.noldus.com) in past work. Other researchers in this domain have used Mangold's INTERACT (www.mangold.de) software (e.g., Fernandez Castelao, Boos, Ringer, Eich, & Russo, 2015; Kolbe et al., 2014). These programs are beneficial, but they still require the judgment and perception of the human coders to both recognize behaviors and assign codes to them.

Researchers often train at least two coders for each study, explaining to them the conceptual and discrete differences among the target behaviors. New coders can be shown video clips from sample recordings in order to practice and calibrate their ability to reliably identify the behaviors of interest. Often, as they become more familiar with the recordings, team context, and process, coders suggest refinements to the coding categories in use. Trained coders work independently, remain blind to any specific hypotheses developed for the study, and meet regularly to compare their coding, discuss discrepancies, arrive at a final agreed to coding for each team's recording, providing the data used to compute necessary intercoder reliability measures. In general, agreement will be higher for discrete events and behaviors (e.g., starting chest compressions) versus verbal statements

(Marsch et al., 2005). This is both because verbal statements often are more difficult to hear (e.g., due to multiple actors talking at the same time) and because verbal statements can be more ambiguous and open to interpretation than behavioral ones. In our experience, most journal reviewers we have encountered in the past seem to expect at least a .70 level of agreement as measured with Cohen's kappa, although many variations on the interrater agreement statistic exist (see Bakeman & Gottman, 1997; LeBreton & Senter, 2008). If one required coders to provide judgments of continuous variables rather than discrete occurrences (e.g., ratings of communication quality or leadership effectiveness), computing indices such as an intraclass correlation, rwg, or the average deviation index would be appropriate (see LeBreton & Senter, 2008).

Our Example

In this particular study, one of the researchers was the primary coder, and a second researcher coded half of the videos. In general, we strongly advise that at least two trained researchers code all videos. Here, though, due to privacy concerns about sending the video files to one another (as we were not co-located during some of the coding phase), we were unable to have two of us code all videos. We did attempt to train another researcher to code the videos, but given that she previously was completely unfamiliar with the context and the terminology used, initial agreement with the primary coder was quite poor. This result further reinforces the point made previously about the necessity of researchers being extremely familiar with the context and the language (e.g., jargon) used.

The coders initially experienced difficulty in coding agreement, and additional calibration efforts were required. The two coders watched several sample video segments together while comparing their coding. Upon doing so, they realized a systematic difference in what they had considered "volitional" communication (see aforementioned). The secondary coder had regarded many more statements as being volitional than did the primary coder, leading to a discrepancy in the number of behaviors coded. Additionally, the coders also realized that a further source of disagreement was difficulty in hearing or understanding specific utterances. When team members spoke in an area of the control room that was not proximal to a microphone, for example, their communications were barely audible. Also, at times, multiple people spoke at the same time, making each actor's statement difficult to hear or decipher. These types of logistical and practical issues are very common and is "the reality" of coding video recordings from the field. While there are ways to address some of these issues (e.g., having each participant wear a microphone), such is not always feasible when collecting data in real organizational contexts. Thus, our advice is obviously to check and do whatever is available to improve the audio and visual quality of the (forthcoming) recordings before data collection begins but also to realize that there likely will still be some specific instances of behavior that remain difficult to code. After taking steps to mitigate these issues, the final interrater reliability achieved was .73.

As noted previously, the primary research objective was determining whether we could distinguish average- versus high-performing teams based on the nature of their interaction patterns. We used two measures of team performance—one was ratings of team effectiveness provided by multiple expert trainers who observed the simulations, and a second was the anticipation ratio, which is a measure of implicit coordination and shared situational awareness (e.g., Entin, Serfaty, & Deckert, 1994).

With respect to the actual coding process, we used Noldus's Observer program in this study. Within this program, one first specifies a list of actors/roles (e.g., supervisor, left board operator, right board operator, etc.) and a list of behaviors (those in Table 2). The program allows one to view both of these lists while also observing the adjacent video on the computer screen. When a relevant behavior occurs in the video recording, the coder clicks on the actor (from the list of actors) and then on the specific behavior exhibited (from the list of behaviors). The program records and timestamps each behavior. Because behaviors sometimes occurred in rapid succession (e.g., when multiple operators noticed and announced a system malfunction simultaneously), we often stopped and rewatched each segment of the video several times to be certain that we heard and appropriately coded each utterance. As such, the coding process can be very time consuming. Here, for example, coding each crew's 15-minute simulation behavior took several hours.

Decision Area 4: Analysis Focus

Overall, decisions concerning analyses using data generated from behavioral coding, similar to most other domains of inquiry, hinge on how best to answer the research question posed and test the hypotheses under consideration. Here, we focus on analysis techniques used in behavioral observation of groups and teams. There are examples of several different analyses using data generated from behavioral observation and coding techniques. For instance, questions surrounding (group comparisons of) frequencies of behaviors, ratings on continuous metrics, or time generally are assessed with chi-square tests, nonparametric tests, analysis of variance, or *t* tests (e.g., Fernandez Castelao et al., 2015; Hunziker et al., 2010). Various types of multiple regression analyses can be used to predict levels of continuous variables (e.g., Westli, Johnsen, Eid, Rasten, & Brattebo, 2010).

Another type of statistical technique common in this domain is lag-sequential analysis, used to identify sequential patterns of coded behavior. In one study, Kauffeld and Meyers (2009) used this procedure to demonstrate that teams engage in solution-oriented and also complaining-oriented sequential patterns. Similarly, Kolbe and colleagues (2014) utilized lag-sequential analysis in showing that higher performing teams engaged in specific types of patterns. Specifically, these more effective teams demonstrated patterns in which "team member monitoring was followed by speaking up, providing assistance, and giving instructions and by patterns in which talking to the room was followed by further talking to the room and not followed by instructions" (Kolbe et al., 2014, p. 1254). Recently, Kennedy and McComb (2014) used neural network techniques—which are especially useful for modeling nonlinear relationships—to study shifts among team process. Examples of studies using these various analyses appear in Table 1.

Some studies involve identifying behavioral differences between lower and higher performing teams, often creating performance clusters of teams (e.g., high performers vs. low performers) based on the performance measures provided by the organization. Comparisons of overall frequencies of coded behaviors between high- and low-performing clusters of teams may be accomplished using *t* tests (e.g., Stachowki et al., 2009; Westli et al., 2010). In addition to comparing frequencies of behaviors, researchers may also choose to investigate differences in the probability of behaviors at certain times. For example, conditional likelihood logit models (see Allison, 1994) may be used in this regard; Waller (1999) used this technique to show that higher performing teams were significantly more likely to engage in certain adaptive behaviors immediately after encountering nonroutine events than were lower performing teams.

Additionally, while lag-sequential analysis can be used to identify sequential patterns of coded behavior, some pattern recognition algorithms such as the THEME algorithm (www.patternvision.com) are able to "ignore" intervening behaviors while identifying a pattern of behaviors that occurs above and beyond chance, depending on the confidence interval and other parameters chosen by the user. Concerning THEME, the algorithm first identifies simple temporal patterns—or "T-patterns"—consisting of two behaviors that sequentially occur significantly more often than by chance. For example, the sequence: "Question (A)—Answer (B)" is a T-pattern, consisting of two behaviors (Question and Answer) that occur in this order more often than by chance. Second, after the significant two-behavior T-patterns are identified, the algorithm cycles through the data hundreds of thousands of times, building more complex hierarchical patterns of relationships among T-patterns. This "bottom-up" approach of pattern detection identifies simple patterns first and then detects larger patterns as a combination of the simpler ones. Third, the algorithm eliminates patterns that are less complete versions of other patterns.

Our Example

Using crew performance data from the plant, we began by comparing the two clusters of teams (high vs. low performing) in terms of the total number of behaviors they exhibited and with respect to each of the specific 11 behaviors. As we have tended to find in most of our studies, frequencies of communication did not differentiate superior performing teams from the other teams. Frequencies, however, only reveal the amount of behaviors, not the patterns of behaviors over time.

To examine whether and how the amount and characteristics of patterns related to team effectiveness, we used the THEME program (described previously). The Observer and THEME programs have interfaces allowing their use in combination. Specifically, the Observer saves a text file of the coded behavior for each crew's simulation, and THEME then analyzes these text files to identify patterns of behavior. THEME then creates a number of indices about any emergent patterns identified. Here, for instance, we examined whether high- versus average-performing crews varied with respect to the amount (i.e., number) of patterns exhibited and with respect to various indices of pattern complexity that THEME produces (e.g., the number of actor switches embedded in the patterns, the number of actors in the patterns, the number of behaviors in the patterns, and the hierarchical complexity of the patterns). Using the THEME results, we compared the two sets of crews on each of these pattern characteristics and also conducted a discriminant function analysis using all five characteristics to distinguish the higher performing crews from the other crews.

We found that the more effective teams engaged in fewer and less complex stable patterns of interaction—findings that we interpreted to suggest the importance of adaptive responses to nonroutine events—events that may not map onto existing schemas and for which standard procedures may not be available. Along with our co-authors, we since have replicated this finding in a study examining trauma teams in high-fidelity simulations (Su et al., 2013). That context differs from the current one significantly with respect to primary team tasks and communication norms. Thus, the fact that we were able to replicate the results in a dissimilar team setting provides initial evidence that less patterned team interaction is more effective across (at least some) domains when teams respond to nonroutine events.

Tools to Enhance Video-Based Approaches

In this final section of the article, we discuss some technological tools that can aid in the coding process and/or lead to new types of insights from video-based data. First, with respect to coding these video recordings, the "Holy Grail" among researchers using this methodology is a program that would conduct automated coding from video, without human coder judgment involved. Such a program not only would save hundreds of hours of work on a given project but also ideally would provide more reliable coding than that which humans can achieve with video data. Communication researcher Joann Keyton and her colleagues (Keyton & DeJoy, 2014; Keyton, Keiser, Graffius, & Primus, 2015) have recently reported measured success in using commercially available products such as dictation software and the game console camera and microphone systems to automate the generation of transcripts from recorded group and team interactions.

In addition to helping automate portions of the coding process, technological advances also can lead to different types of questions and insights with video data. For example, by using programs such as Linguistic Inquiry and Word Count (LIWC; http://liwc.wpengine.com), one can add content coding to behavioral observation coding (Pennebaker, Francis, & Booth, 2001). Given that, as noted previously, the mere frequency of communication often fails to correspond to team effectiveness, investigating both the temporal pattern and the content of communication can be more a fruitful approach in behavioral observation approaches. Furthermore, researchers may add the additional layer of various paralingual aspects of communication using video recorded data. For instance, PRAAT, a free downloadable paralanguage system, analyzes features of speech such as pitch, intensity, and voice breaks (http://www.fon.hum.uva.nl/praat/). The number of research areas and questions for which voice characteristics would be of interest likely is great and varied. A simple application of this approach would be examining how characteristics of leader speech relate to team member behavior, affect, motivation, and so on—perhaps beyond the amount and content of what the leader says. Another program that seems quite promising is FaceReader from Noldus (www.noldus.com). With some limitations related to the angle of the face visible on video, this program can be used to produce a continuous measure of a number of basic emotions as portrayed by video recorded facial expressions. As research suggests that the configuration of positive affect levels across team members influences team effectiveness in crisis situations (for example, Kaplan, LaPort, & Waller, 2013), using tools such as FaceReader would enable researchers to add an affective component to video-based behavioral analyses.

This is just a sampling of the programs that one could use to complement the type of coding we have described here. Others certainly exist, and we imagine that future technological advances will result in many more programs in the coming years. Researchers can use these tools to address novel questions or address existing questions in a novel manner. Furthermore, exploiting the video recorded data by using several of these programs in concert may allow for investigating especially interesting questions and providing particularly enlightening conclusions. For instance, extending the aforementioned leader example, in addition to analyzing leaders' voice characteristics, one also could code the content of the leaders' communications as well as leaders' emotional expressions when delivering them. Doing so, researchers could investigate questions such as whether these various factors operate in an additive versus compensatory manner across different situations in impacting team member affect, motivation, and similar outcomes.

Conclusion

The study of behavior in modern organizations largely began with the careful analysis of human movement using data derived from direct observation or film. While the nature of work fundamentally has changed in the past century, (the study of) organizational behavior is still about *behavior* (Campbell, 1990). Using systematic methodologies to code video-based data and increasingly sophisticated statistical programs to analyze the resultant coding can lead to innovative questions and enlightening conclusions that more traditional data sources (e.g., surveys) simply cannot provide. Thus, we see the application of quantitative methods for video recorded data to still be highly relevant and useful. We hope to have provided researchers with some useful guidance to aid in implementing this approach.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Allison, P. D. (1994). Using panel data to estimate the effects of events. *Sociological Methods and Research*, *23*, 174-199.

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge, UK: Cambridge University Press.

Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. New York, NY: Cambridge University Press.

Bales, R. F. (1950). *Interaction process analysis: A method for the study of small groups*. Oxford, UK: Addison-Wesley Press.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology (Vol. 1*, 2nd ed., pp. 687-731). Palo Alto, CA: Consulting Psychologists Press.

DeVita, M. A., Schaefer, J., Lutz, J., Wang, H., & Dongilli, T. (2005). Improving medical emergency team (MET) performance using a novel curriculum and a computerized human patient simulator. *Quality and Safety in Health Care*, *14*, 326-331.

Entin, E. E., Serfaty, D., & Deckert, J. C. (1994). *Team adaptation and coordination training (TR-648–1)*. Burlington, MA: Alphatech.

Fernandez Castelao, E., Boos, M., Ringer, C., Eich, C., & Russo, S. G. (2015). Effect of CRM team leader training on team performance and leadership behavior in simulated cardiac arrest scenarios: A prospective, randomized, controlled study. *BMC Medical Education*, *15*, 116.

Fernandez Castelao, E., Russo, S. G., Cremer, S., Strack, M., Kaminski, L., Eich, C., … Boos, M. (2011). Positive impact of crisis resource management training on no-flow time during simulated cardiopulmonary resuscitation: a closer look at verbal coordination behaviour. *Resuscitation*, *82*, 1338-1343.

Homans, G. C. (1951). *The human group*. London: Routledge and K. Paul.

Hunziker, S., Buhlmann, C., Tschan, F., Balestra, G., Legeret, C., Schumacher, C., … Marsch, S. (2010). Brief leadership instructions improve cardiopulmonary resuscitation in a high fidelity simulation: A randomized controlled trial. *Critical Care Medicine*, *38*, 1086-1091.

Israel, M. (2014). *Research ethics and integrity for social scientists: Beyond regulatory compliance* (2nd ed.). London: Sage Publications Ltd.

Kaplan, S. A., LaPort, K., & Waller, M. J. (2013). The role of positive affect in team performance during crises. *Journal of Organizational Behavior*, *34*, 473-491.

Kauffeld, S., & Lehmann-Willenbrock, N. (2012). Meetings matter: Effects of team meetings on team and organizational success. *Small Group Research*, *43*, 130-158.

Kauffeld, S., & Meyers, R. A. (2009). Complaint and solution-oriented circles: Interaction patterns in work group discussions. *European Journal of Work and Organizational Psychology*, *18*, 267-294.

Kennedy, D. M., & McComb, S. A. (2014). When teams shift among processes: Insights from simulation and optimization. *Journal of Applied Psychology*, *99*, 784-815.

Keyton, J., & DeJoy, D. (2014). *Moving group talk to transcript*. Raleigh, NC: Interdisciplinary Network for Group Research.

Keyton, J., Keiser, J. M., Graffius, J. M., & Primus, X. J. (2015). *Kinect for group research*. Pittsburgh, PA: Interdisciplinary Network for Group Research.

Kolbe, M., & Boos, M. (2009). Facilitating group decision-making: Facilitator's subjective theories on group coordination. *Forum: Qualitative Social Research*, *10*, Art.28.

Kolbe, M., Burtscher, M., & Manser, T. (2013). Co-ACT—A framework for observing coordination behavior in acute care teams. *BMJ Quality & Safety*, *22*, 596-605.

Kolbe, M., Grote, G., Waller, M. J., Wacker, J., Grande, B., Burtscher, M., … Spahn, D. R. (2014). Monitoring and talking to the room: Autochthonous coordination patterns in team interaction and performance. *Journal of Applied Psychology*, *99*, 1254-1267.

Kolbe, M., Strack, M., Stein, A., & Boos, M. (2011). Observing coordination in human group decision-making: MICRO-CO—A micro-analytical taxonomy for analysis of coordination mechanism in decision-making groups. In M. Boos, M. Kolbe, P. Kappeler, & T. Ellwart (Eds.), *Coordination in human and primate groups* (pp. 199-219). Heidelberg: Springer.

Kozlowski, S. W. J., & Bell, B. S. (2003). Work groups and teams in organizations. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology: Industrial and organizational psychology* (Vol *12*, pp. 333-375). London: Wiley.

Lancaster, J. (2004). *Making time: Lillian Moller Gilbreth, a life beyond Cheaper by the Dozen*. Boston, MA: Northeastern University Press.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815-852.

Lehmann-Willenbrock, N., Meinecke, A. L., Rowold, J., & Kauffeld, S. (2015). How transformational leadership works during team interactions: A behavioral process analysis. *Leadership Quarterly*, *26*, 1017-1033.

Lei, Z., Waller, M. J., Hagen, J., & Kaplan, S. (2015). Team adaptiveness in dynamic contexts: Contextualizing the roles of interaction patterns and in-process planning. *Group & Organization Management*. Advance online publication. doi:10.1177/1059601115615246

Lim, S., & Murnighan, J. K. (1994). Phases, deadlines, and the bargaining process. *Organizational Behavior and Human Decision Processes*, *58*, 153-171.

Marsch, S. C., Tschan, F., Semmer, N., Spychiger, M., Breuer, M., & Hunziker, P. R. (2005). Performance of first responders in simulated cardiac arrests. *Critical Care in Medicine*, *33*, 963-967.

Nelson, D. (1980). *Frederick W. Taylor and the rise of Scientific Management*. Madison, WI: University of Wisconsin Press.

Noldus, L. P. J. J., Trienes, R. J. H., Hendriksen, A. H. M., Jansen, H., & Jansen, R. G. (2000). The Observer Video-Pro: New software for the collection, management, and presentation of time-structured data from videotapes and digital media files. *Behavior Research Methods, Instruments, & Computers*, *32*, 197-206.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC2001*. Mahwah, NJ: Lawrence Erlbaum Associates.

Stachowski, A., Kaplan, S. A., & Waller, M. J. (2009). The benefits of flexible team interaction during crises. *Journal of Applied Psychology*, *94*, 1536-1543.

Stanton, N. (1996). *Human factors in nuclear safety*. London: Taylor & Francis.

Stoolmiller, M., Eddy, J., & Reid, J. B. (2000). Detecting and describing preventive intervention effects in a universal school-based randomized trial targeting delinquent and violent behavior. *Journal of Consulting and Clinical Psychology*, *68*, 296-306.

Su, L., Kaplan, S., Burd, R., Hargrove, A., Winslow, C., & Waller, M. (2013). The best trauma teams demonstrate teamwork patterns similar to the best nuclear power plant teams. *Critical Care in Medicine*, *41*, A123.

Su, L., Waller, M. J., Kaplan, S., Watson, A., Jones, M., & Wessel, D. L. (2015). Cardiac resuscitation events: One eyewitness is not enough. *Pediatric Critical Care Medicine*, *16*, 335-342.

Waller, M. J. (1999). The timing of adaptive group responses to non-routine events. *Academy of Management Journal*, *42*, 127-137.

Weick, K. (1985). Systematic observation methods. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology. Vol. 1, Theory and method* (pp. 567-634). New York, NY: Random House.

Weingart, L. R. (1997). How did they do that? The ways and means of studying group process. *Research in Organizational Behavior*, *19*, 189-239.

Westli, H. K., Johnsen, B. H., Eid, J., Rasten, I., & Brattebo, G. (2010). Teamwork skills, shared mental models, and performance in simulated trauma teams: An independent group design. *Scandinavian Journal of Trauma Resuscitation, and Emergency Medicine*, *18*, 47.

Xiao, Y., Hunter, W. A., Mackenzie, C. F., Jefferies, N. J., & Horst, R. L. (1996). Task complexity in emergency medical care and its implication for team coordination. *Human Factors*, *38*, 636-645.

Author Biographies

Mary J. Waller, PhD (University of Texas at Austin), is Professor of Organizational Studies at the Schulich School of Business, York University, Toronto. For more than two decades, she has focused her research efforts on understanding team dynamics and effectiveness during critical situations. Her work appears in management, psychology, and health care publications including *Academy of Management Journal, Academy of Management Review, Management Science, Journal of Organizational Behavior, Journal of Applied Psychology,* and *Pediatric Critical Care Medicine*. She has served on the Board of Governors of the Academy of Management, and currently serves on the editorial boards of *Academy of Management Discoveries and Organization Science*.

Seth A. Kaplan, PhD (Tulane University) is an Associate Professor of Industrial/Organizational (IO) Psychology at George Mason University, Fairfax, Virginia. His research focuses on understanding the determinants of team effectiveness in high reliability and extreme contexts. His scholarly work has appeared in journals including *Psychological Bulletin, Journal of Applied Psychology, Journal of Management*, and *Journal of Organizational Behavior*. He currently serves on the editorial boards of *Journal of Applied Psychology, Organizational Research Methods, Journal of Business* and *Psychology,* and *Journal of Management*.