Using live disaster exercises to study large multiteam systems in extreme environments: Methodological and measurement fit

Sara Waring
University of Liverpool, UK

Abstract
Multiteam systems (MTSs) are comprised of two or more teams working toward shared subordinate goals but with unique subgoals. In large MTSs operating in extreme environments, coordination difficulties have repeatedly been found, which compromise response effectiveness. Research is needed that examines MTSs in situ within extreme environments to develop temporal theories of inter-team processes and understanding of how coordination may be improved within these challenging contexts. Live disaster exercises replicate the complexities of extreme environments, providing a valuable avenue for observing inter-team processes in situ. This article seeks to contribute to MTS research by highlighting (i) a mixed-method framework for collecting data during live disaster exercises that uses both inductive and deductive approaches to promote methodological and measurement fit; (ii) ways in which data can be collected and combined to meet the appropriate standards of their methodological class; and (iii) a case example of a National exercise.

Keywords
multiteam system, live exercises, disaster response, methodological fit, measurement fit

Traditional theory-driven research primarily seeks to advance and refine theory through testing theoretically derived hypotheses, with real-world applications serving a secondary goal. While this is undoubtedly important, becoming too narrowly focus on filling gaps in...
existing theory can lead to academic insights that are so far removed from organizational contexts they fail to provide any real-world value (Campbell, 1990; Schwarz & Stensaker, 2014). It may also prevent researchers from reporting "rich phenomena for which no theory yet exists," which is an important precursor for developing new constructs (Hambrick, 2007, p. 1346). In contrast, phenomenon-driven research is based on abductive inference (Meyer & Lunnay, 2013), focusing on describing, documenting, and conceptualizing a real-world problem, and leveraging and modifying existing theory, or developing new theory to better understand and address it (Mathieu, 2016; Schwarz & Stensaker, 2014). Phenomenon-driven research can uncover new ideas, concepts, and relationships that may subsequently be tested, along with establishing the veracity and conditions under which existing theory holds (Mathieu, 2016; Shuffler & Carter, 2018).

One example of a complex real-world problem in need of focus is the coordination difficulties that occur during disaster response and other extreme environments characterized by risk, uncertainty, and need for rapid action. The multiteam systems (MTSs) that form to respond to disasters are large, comprised of several teams working toward shared superordinate goals but with unique subgoals at individual and team levels (Marks, Mathieu, & Zaccaro, 2001). Membership is determined by goal and task interdependencies that span several organizations, including police, fire, ambulance, and health, creating a diverse pool of knowledge and resources (Marks, DeChurch, Mathieu, Panzer, & Alonso, 2005). However, coordination difficulties are repeatedly identified in MTSs responding to disasters and other extreme environments, with potentially severe consequences for public safety (Bharosa, Lee, & Janssen, 2010; DeConstanza, DiRosa, Jimenez-Rodriguez, & Cicin-Sain, 2014; Kerslake, 2018; Majchrzak, Jarvenpaa, & Hollingshead, 2007; Marks et al., 2001; Patrick, 2011; Pollock, 2013; Waring et al., 2018). Studying this phenomenon is important for identifying new theoretical constructs to improve understanding of MTS functioning in extreme environments and provide an evidence base to inform practice.

However, while the growth in MTS research over the past two decades has provided valuable insights into inter-team processes, questions have been raised regarding the extent to which findings apply to large MTSs operating in extreme environments (Shuffler, Jimenez-Rodriguez, & Kramer, 2015). Firstly, this body of research predominantly uses controlled experimental methods to study small MTSs comprised of two or three component teams with narrow specializations and goals, completing computer-generated tasks (Bienefeld & Grote, 2013; Carter, 2014; Cobb, 1999; Davison, Hollenbeck, Barnes, Sleesman, & Ilgen, 2012; Firth, Hollenbeck, Miles, Ilgen, & Barnes, 2015; Marks et al., 2005). However, the MTSs that respond to disasters are much more complex and varied in size, shape (compatibility and separation of goals, knowledge, working practice, and capabilities between component teams), and dynamism (variability and instability of the system over time) (Luciano, DeChurch, & Mathieu, 2015). The contexts that these MTSs operate in are also far more complex, diverse, fast paced, time pressured, risky, and uncertain than experimental studies have captured (Waring et al., 2018). Paying greater attention to where MTSs live and operate and observing MTSs in their natural settings will provide important opportunities for future research (Shuffler & Carter, 2018).

In addition, the way in which MTSs have been studied to date also provides limited insights into temporal aspects relating to how and why MTS phenomena emerge and change over time and what factors affect this (Luciano et al., 2015; Shuffler & Carter, 2018). Observations of inter-team behaviors have usually been conducted in relation to controlled tasks of a short duration, which result in inter-team processes being treated as static phenomena (Kozlowski, 2015; Shuffler & Carter, 2018).
Even field studies that examine large MTSs operating in extreme environments (de Vreede, Briggs, & Reiter-Palmon, 2010) have largely been restricted to post-incident accounts and self-report questionnaires rather than to methods that capture dynamic changes in inter-team processes due to risks posed to researcher safety (Crowe, Allen, & Bowes, 2014; DeChurch et al., 2011; DiRosa, 2013; O’Sullivan, 2003) (see Table 1 for summary). Longitudinal research that captures growth trajectories and fluctuations over time is needed to promote the dynamic nature of team research and develop nascent temporal theories of inter-team processes.

The following paper focuses on a promising option for conducting research to develop temporal theories of inter-team processes and improve understanding of MTS functioning in extreme environments. Live disaster exercises physically and psychologically replicate the complexities of extreme events but with minimal risks to safety (Healey, Hodgkinson, & Teo, 2009; Waring et al., 2018). This presents opportunities to adopt a range of qualitative and quantitative methods to examine behaviors, cognitions and affective states in situ to capture growth trajectories and fluctuations. However, using live disaster exercises to conduct research is not without its challenges in terms of ensuring the credibility of findings. Accordingly, this article presents a data collection framework that is both multi-method (multiple sources of qualitative data) and mixed-method (qualitative and quantitative) and adopts inductive and deductive approaches to promote methodological (Edmondson & McManus, 2007) and measurement fit (Luciano, Mathieu, Parks, & Tannenbaum, 2018) (see Figure 1 for an overview of the data collection framework). As will be detailed, inductive approaches are used to identify and define key constructs such as how MTS phenomena manifest, under what conditions and why, to develop nascent theory (Edmondson & McManus, 2007; Eisenhardt, Graebner, & Sonenshein, 2016; Kozlowski, 2015) and tools and methods needed for deductive approaches to test temporal frameworks (Luciano et al., 2018).

In summary, drawing on over a decade of experience working with law enforcement and emergency services, the following paper details a research approach for developing nascent theories of inter-team processes in MTSs operating in extremis using live disaster exercises. In particular, the paper highlights (i) a data collection framework that promotes methodological and measurement fit; (ii) ways in which data can be collected and combined to meet the appropriate standards of their methodological class; and (iii) a case example of a National exercise. Rather than repeating discussions documented in-depth elsewhere regarding qualitative and quantitative data analysis (Lyons & Coye, 2016; Marvasti, 2014; Thorne, 2000), this article provides a set of key considerations for reviewers and researchers alike when assessing data collection in live disaster exercises.

Methodological and measurement fit in live disaster exercises

Live disaster exercises provide a novel and appropriate way of collecting contextually rich data from practitioners with responsibility for managing real disasters (Smith, Dowell, & Ortega-Lafuente, 1999), which is beneficial for examining the impact of experience, knowledge, organizational culture, and policies on inter-team processes. What makes these exercises particularly beneficial, both for practitioners testing their preparedness and researchers studying inter-team processes in situ, is their ability to elicit similar cognitive and emotional responses (psychological fidelity) to those evoked in real contexts (Berlin & Carlström, 2015). This is achieved using realistic goals, choices, and decisions and hiding certain elements from players beforehand (Brehmer & Dorner, 1993). These exercises also promote physiological immersion by replicating physical
Table 1. Features of studies of multiteam systems.

| Authors                  | Core research questions                                                                 | Sample                                                                 | Methods                                                                 |
|--------------------------|------------------------------------------------------------------------------------------|------------------------------------------------------------------------|------------------------------------------------------------------------|
| Bienefeld and Grote (2013) | How does shared leadership within and between teams in MTSs affect team goal attainment and MTS success? | 84 MTSs of cockpit and cabin crews, each one consisting of three dyads ($N = 504$) | Observation during standardized 30-min simulations of in-flight emergencies to identify the impact of centralized versus shared leadership on the ability of MTSs to achieve goals. |
| Carter (2014)            | How does leadership network structure impact innovation within MTSs?                       | 49 MTSs, each one consisting of three or four teams of between two and four undergraduate students ($N = 456$) | Self-report questionnaires measuring leadership structure, concentration, mutuality, and within- and between-team density, completed after collaboration on a semester long innovation-focused group project. |
| Cobb (1999)              | How do environmental complexity and team training impact team processes and performance in MTSs? | 36 MTSs, each one consisting of four undergraduate students split into two dyads ($N = 144$) | Observation during low fidelity computer-generated combat flight simulations, and interviews, to identify differences in MTS performance between teams who had and had not received team coordination training in low- and high-complexity environments. |
| Crowe, Allen, and Bowes (2014) | How can multi-crew responses be changed to improve the efficiency and safety of disaster management? | Over 100 firefighters were involved in the incident response but details of the MTS structure are not provided | Case study analysis based on information contained within an official post-incident report relating to a fire in which a firefighter lost his life to identify what factors hindered performance and safety and why. |
| Davison, Hollenbeck, Barnes, Sleesman, and Ilgen (2012) | How does horizontal and vertical coordination impact MTS performance within and between component teams? | 233 MTSs, each one consisting of 14-person teams of Air Force captains split into two six-member component teams and a six-member integration team ($N = 3,262$) | Observation of horizontal and vertical coordination during repeated episodes of computer-simulated combat missions, each episode lasting approximately 2 hr and consisting of 10 scenario rounds. |
| Authors | Core research questions | Sample | Methods |
|---------|-------------------------|--------|---------|
| DeChurch et al. (2011) | What behaviors and functions do leaders of MTSs adopt when operating in extreme contexts? | 110 post-incident accounts of critical incidents—details of the MTS structures are not provided | Historiometric analysis of post-incident accounts using critical incident and grounded theory techniques to identify what leadership behaviors were exhibited and the function of leadership across incidents. |
| de Vreede, Briggs, and Reiter-Palmon (2010) | How do parallel and serial modes of organizing group processes affect brainstorming in MTSs? | Two MTSs, one comprised of 100 members from multiple organizations, the other comprised of 500 members of a single organization. Component teams consisted of between 4 and 11 participants—details of the number of component teams within each MTS are not provided | Field study consisting of observations of group meetings to measure productivity in terms of number of ideas and number of unique ideas generated, level of elaboration and redundant comments, in addition to post-incident self-report questionnaires measuring satisfaction with meeting process and outcome. |
| De Vries, Hollenbeck, Davison, Walter, and Van Der Vegt (2016) | What is the relationship between vertical and horizontal coordination, interpersonal functional diversity, and multiteam system performance? | 236 MTSs, each one consisting of 14 US Air Force Officers split into two six-member component teams comprised of four operational staff and two boundary spanners, plus a six-member integration team comprised of the four boundary spanners and two additional members ($N = 3,304$) | Observations of horizontal and vertical coordination conducted during repeated episodes of computer-generated combat missions, each episode lasting approximately 2 hr and consisting of 10 scenario rounds. |
| DiRosa (2013) | How does the development of cohesion within and between component teams affect MTS combat readiness? | 637 U.S. Army soldiers from across 128 squads belonging to 38 platoons—details of number of component teams and members within each MTS are not provided | Field study consisting of self-report questionnaires measuring between-squad interdependence, leader boundary spanning, goal alignment, within- and between-squad cohesion, and platoon readiness. |
| Firth, Hollenbeck, Miles, Ilgen, | How can representational gaps be reduced and what is the impact of this on | 249 MTSs, each one consisting of 14 U.S. Air Force captains split into | Observations conducted during repeated episodes of computer-simulated |
| Authors                        | Core research questions                                      | Sample                                                                 | Methods                                                                                                                                 |
|-------------------------------|--------------------------------------------------------------|------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------|
| and Barnes (2015)             | between-team coordination and MTS performance?               | two six-member component teams and one six-member leadership team (N = 3,486) | combat, each lasting approximately 2 hr and consisting of 10 scenario rounds, to measure differences in MTSs who had and had not received frame-of-reference training on within- and between-team coordination and MTS performance |
| Lanaj, Hollenbeck, Ilgen, Barnes, and Harmon (2013) | How does decentralized planning impact multiteam system performance? | 210 MTSs, each one consisting of 14 U.S. Air Force Officers split into two six-member component teams comprised of four operational staff and two boundary spanners, plus a six-member integration team comprised of the four boundary spanners and two additional members (N = 2,940) | Observations of horizontal and vertical coordination and risk strategies conducted during repeated episodes of computer-simulated combat, each lasting approximately 3 hr (number of scenario rounds was not reported) to measure differences in MTSs who received pre-simulation centralized planning sessions compared to decentralized planning sessions. |
| Marks, DeChurch, Mathieu, Panzer, and Alonso (2005) | How do component teams within an MTS integrate their efforts to collectively succeed? | 46 MTSs, each one consisting of four undergraduate students split into two dyads (N = 184) | Observations conducted during three repeated episodes of computer-generated flight simulations, each lasting 15-min, and post-incident interviews, to measure transition and action phase processes, and MTS performance |
| O'Sullivan (2003)              | How does standardization and synchronization affect information and workflow within a multiagency MTS operating virtually? | Research was conducted in relation to an aerospace product-development project involving a lead firm and 20 supplier organizations—details of participant numbers and MTS structures are unclear | Field study consisting of 78 semi-structures interviews, observations of 160 meetings and, reviewing technical and administrative documents to examine the division of labor across geographically dispersed teams across design phases, and the impact of this on performance. |
Figure 1. Data collection framework showing how methods influenced one another throughout data collection. *Questions asked during post-exercise debriefs were informed by data collected during phase one and phase two and provided an additional source of data for triangulating qualitative findings to improve trustworthiness.
features of the real world using live actors and real equipment (Cohen et al., 2012). This allows the impact of subtle or unexpected aspects of environments on human processes to be examined, such as geographic distance, equipment, visual cues, exhaustion, and concurrently managing cognitive and manual tasks (Issenberg & Scalese, 2008).

In addition, as live disaster exercises run for a period of several hours to several days, they provide a valuable means of studying temporal relationships between inter-team processes and addressing corresponding questions that remain outstanding. For example, little is known about the enabling conditions that positively influence MTS functioning in extreme environments and how these conditions develop over time (Shuffler & Carter, 2018). Similarly, while research shows that vertical coordinated action (VCA) between specialized task and system-wide integration teams can improve MTS performance (Davison et al., 2012; De Vries, Hollenbeck, Davison, Walter, & Van Der Vegt, 2016), little is known about the antecedents that promote VCA. Live disaster exercises therefore have the potential to make valuable theoretical contributions regarding how, when, and why inter-team processes emerge, provided that careful consideration is given to methodological fit and measurement fit.

**Methodological fit** refers to the degree of “internal consistency among elements of a research project” and emphasizes the importance of matching research approach and method with the state of theoretical development (Edmondson & McManus, 2007, p. 1155). Whereas nascent theory is best suited to inductive approaches and open-ended qualitative methods that seek to explore novel contexts and phenomena in depth, mature theory possesses the well-developed constructs, models, and knowledge bases needed for deductive hypothesis testing approaches and focused quantitative measures. Intermediate theory in which explanations for phenomena are provisional is better suited to a hybrid approach that draws on a blend of qualitative data to elaborate a phenomenon and quantitative data to provide preliminary tests of relationships and promote important insights (Luciano et al., 2018; Yauch & Steudel, 2003). This process of moving between inductive theory creation and deductive theory-testing processes has long been advocated as beneficial to theory development (Cialdini, 1980; Edmondson & McManus, 2007; Fine & Elsbach, 2000; Weick, 1979).

While MTS theory is in the intermediate stage, understanding of dynamic temporal relationships over time remains nascent, as does understanding of inter-team processes in extreme environments. To test temporal aspects of MTS theory, knowledge of what constructs to measure and how frequently must first be developed, along with appropriate tools for measuring constructs within the context of study. Initially adopting inductive followed by deductive approaches will provide a better methodological fit that allows temporal relationships to be preliminarily tested. A framework for moving from inductive to deductive stages of inquiry within a live disaster exercise is presented in the case study section to demonstrate how this can be achieved (see Figure 1 for an overview of these stages).

**Measurement fit** refers to the degree of alignment between how a construct is conceptualized and examined (Luciano et al., 2018). While the terminology is problematic for inductive approaches that explicitly reject “measurement,” the concept does highlight important implications for improving the degree to which inferences can legitimately be made from operationalization in a study to the phenomenon of interest (Trochim, 2006). There is no one-size-fits-all approach for achieving measurement fit (Luciano et al., 2018). For example, the number of telephone calls made between team members may provide an index of knowledge sharing in geographically dispersed teams but is less appropriate in physically collocated teams. Instead, achieving measurement fit is an iterative process that
requires consideration of (i) construct elements, (ii) measurement features, and (iii) contextual considerations (Luciano et al., 2018). Each component poses important implications for making methodological choices to study MTSs using live disaster exercises.

(i) **Construct elements** are concerned with clearly defining the construct space to ensure good fit between conceptualization and measurement. This includes defining inclusion and exclusion rules to clarify the underlying dimensions of the construct and the nature of the domain it applies to such as the people, events, or objects (Podsakoff, MacKenzie, & Podsakoff, 2016). It also includes defining the appearance of the construct in terms of how it is likely to manifest, what conditions are required for the manifestation to occur, and how the shape of the construct may change over time to provide a better understanding of how phenomena arise, continue, and cease (Luciano et al., 2015). Achieving a well-defined construct space is vital for ensuring clarity in what is being measured and why.

However, for nascent aspects of theory, such as temporal relationships in inter-team processes and understanding of MTS functioning in extremis, a construct space does not yet exist and must first be developed from the ground up. Initially, inductively driven qualitative methods are needed to give primacy to first-hand lived experiences to define the underlying dimensions of the construct and the domain that it applies to (phase one), followed by deductive approaches (phase two) to measure and test phenomenon (Corley & Gioia, 2011). Live disaster exercises provide a contextually rich setting to generate the detailed descriptions needed of what inter-team processes occur in situ, when and how to develop a construct space for MTS functioning in extreme environments. Such knowledge can then be used to highlight what features are important to measure and design tools to do so.

(ii) **Measurement features** consider how data are collected and operationalized, including what content is contained in a measure to ensure that it aligns with the construct space and adequately samples and represents the construct domain at the appropriate level of specificity (content validity). It also includes aligning the measurement technique to the construct elements to ensure that data collection captures the intended phenomena (construct validity). The temporal theory guiding an investigation is important to this, ensuring that data collected accurately capture the emergence and changes in intended phenomena over time. Examining temporal relationships requires longitudinal approaches to capture constructs repeatedly over time (Luciano et al., 2018), articulating relationships and their positioning within a relevant temporal framework (Cronin, Weingart, & Todorova, 2011).

Much of the extant MTS research has either used controlled environments and tasks lasting a matter of minutes or field studies that rely on post-incident measures (see Table 1 for summary). It is self-evident that temporal factors cannot be explored using static variables or measurement techniques. For example, the as-yet-unknown construct shape of MTS coordination in extreme environments cannot be determined a priori. Live disaster exercises provide an opportunity to collect the data needed to inform temporal frameworks, such as detailed descriptions of when and why phenomena occur and change (phase one), which may subsequently support decisions regarding.
when to measure phenomena across time points (phase two). Methodological fit goes to the core of the issue as methods that are inappropriate for the task and state of theory development compromise findings.

Measurement features also include considering the source that data are collected from, ensuring that these individuals have sufficient knowledge to provide valuable information and are motivated to give accurate responses that are grounded in the study context and construct elements. Live disaster exercises are beneficial for researchers interested in studying temporal relationships in inter-team processes in extremis because they are responded to by the same practitioners as real disasters. In contrast to naive participants, they possess knowledge of organizational policies, guidance, principles, and constraints. Prolonged engagement within this study context can also build trust and rapport with practitioners, encouraging them to provide the accurate, detailed, rich responses needed to demonstrate trustworthiness.

In addition, measurement features are concerned with aggregation of data from streams, individuals, sources, or times to provide a meaningful sample of behavior, cognition, or affect that gives a valid snapshot of the phenomena. For quantitative data, justification for aggregation comes from use of interclass correlations, scale reliability, and interrater agreements (LeBreton & Senter, 2008). For qualitative researchers, data triangulation and drawing on multiple sources of data to form conclusions are important for enhancing trustworthiness (Casey & Murphy, 2009). Here, the issue is not whether results are replicable across multiple experiments, but whether multiple sources of data from the same context, each strong enough to stand on its own merits, converge to identify a common set of findings that explain what is happening within these contexts and why (Ormerod & Ball, 2010).

(iii) Contextual considerations are also important for judging the appropriateness of research methodologies. As mentioned above, methodological fit is one such consideration (Edmondson & McManus, 2007). Indeed, the decision to use live disaster exercises is influenced by the nascent nature of temporal theory of inter-team processes in extremis and need to observe processes in situ. Live disaster exercises present many of the contextual complexities of real disasters over timescales of several hours or days, but with reduced risks to safety, providing opportunities to study temporal aspects of phenomena in situ.

Study context also poses implications for whether and how phenomena of interest can be examined (Luciano et al., 2018). This relates to what Lipshitz (2010) calls “substantive rigor,” the extent to which methods offer the best potential for obtaining trustworthy answers to research questions within the constraints of the context (Lipshitz, 2010). Methods that disrupt the realism of practitioner responses within any field study, including live disaster exercises, reduce how credibly data capture cognition, emotion, and behavior in situ (Crandall, Klein, & Hoffman, 2006). As with any field research, live disaster exercises may also have logistical restrictions in terms of level of access and intrusion the host setting and participants will tolerate, how much time they are willing to give, and how frequently an event of interest occurs (Luciano et al., 2018). Intrusive methods are likely to raise objections from exercise planners, as their main priority is to ensure exercises test their organizational responses as realistically as possible (Crandall et al., 2006).

From experience, engaging with exercise planning teams throughout the process is beneficial for gaining a better understanding of the purpose and logistics of how the exercise will run. It also provides opportunities to consult with these subject matter experts (SMEs)
regarding data collection methods to improve measurement fit by identifying less intrusive ways of accessing similar data. For example, while getting close enough to hear verbal communications may affect the realism of responder interactions, consultation with SMEs can identify alternative solutions such as accessing footage from body cameras that responders are sometimes already wearing during an exercise. Indeed, as will be demonstrated in the case example, working alongside agencies during exercise-planning poses many benefits for improving alignment of methods to research questions within the context of live disaster exercises and adhering to the rules of that methodological class to improve the trustworthiness of findings.

Validity and trustworthiness in live disaster exercises

While live disaster exercises present opportunities to study inter-team processes in situ over time using both quantitative and qualitative methods, conducting research in these contexts can pose challenges for meeting scientific standards (Goldthorpe, 2000; Joppe, 2000; Popper, 2002). Some researchers argue that such standards should be relaxed for particularly interesting or important theories (Sutton & Staw, 1995). Others argue that demonstrating credibility is vital for findings to be worthy of attention (Golafshani, 2003; Lincoln & Guba, 1985). In line with the latter view, rather than seeking to relax standards, this article highlights ways of collecting and combining data within live disaster exercises to improve credibility. The following brief overview serves to clarify why combining data is important for meeting the standards of each methodological class.

Quantitative methods of data collection seek to objectively measure key concepts in part to demonstrate validity. One key aspect of this is construct validity, the degree to which a test measures what it claims to measure. This is often demonstrated by comparing a test to others that measure similar constructs to check they correlate (Trochim, 2006). However, given the somewhat nascent state of temporal inter-team processes and MTS functioning in extremis, researchers are still searching for key constructs, let alone having tests to measure them and make comparisons. Where nascent theory development leads to the creation of new tests, this raises the issue of how best to demonstrate construct validity in the absence of other tests for comparison.

One way in which this is done in field research is by adopting mixed methods to integrate and triangulate qualitative and quantitative findings. Contextually rich qualitative data are used to build a detailed understanding of the construct, which is important for showing how the new measure relates to this (Edmondson & McManus, 2007). For example, qualitative data sources such as interviews with SMEs and practitioners may be used to develop an observational coding framework to measure frequency of behaviors associated with shared knowledge of roles and responsibilities in live disaster exercises (Waring et al., 2018). However, alongside counting behavioral frequencies, it would also be beneficial to keep detailed observational descriptions to improve understanding of the context in which behaviors occur and whether there are patterns in what happens as a result of these behaviors. In-depth interviews with practitioners would provide further contextual detail regarding when, why, and how such behaviors are used and the impact of this.

Another key standard for quantitative methods is internal validity, confidence in the inferences drawn about cause and effect relationships. Internal validity is concerned with demonstrating that variations in a dependent variable result from variations in the independent variable(s) rather than from other confounding variables (Abernethy, Chua, Luckett, & Selto, 1999). This is usually achieved by controlling extraneous variables, standardizing
procedures and measures (Black, 1999). As with other field studies, procedures and measures may be applied in a standardized way but the novel, dynamic nature of disaster exercises prevents exact replication and control over extraneous variables, making causality difficult to determine (Koch & Harrington, 1998). It is therefore vital to monitor and document the context in which quantitative tests are administered to make accurate causal conclusions (McMillan, 2007). For example, during a live disaster exercise, measures of performance ratings against standardized criteria may be taken at regular intervals to capture changes. Qualitative descriptions of context and behaviors observed, a rationale for each rating, and interviews and self-report scales completed by practitioners would all assist with drawing conclusions about the causes and impact of changes in performance. Combining data sources in this way also allows for potential weaknesses in one source to be compensated by another. For example, qualitative methods being used to account for potential confounds when administering surveys.

Another key standard for quantitative methods is external validity, generalizing relationships found within a study to other people, times, or contexts. The specialized nature of practitioners engaged in live disaster exercises and other extreme environments limits ability to adopt strategies relating to the population validity aspect of external validity such as randomization, random sampling, and large sample sizes (Drost, 2011). In contrast, these live exercises replicate the complexity of disasters, improving the ecological validity of research and generalizability of findings to real disaster response (Lipshitz, 2010). Balancing internal and external validity is notoriously difficult to do because environments cannot be both controlled and demonstrate real-world complexity (Drost, 2011). This does not mean that research using live disaster exercises should only be judged against the criteria of ecological validity or that high ecological validity compensates for lower internal and population validity. It is important to adopt mixed methods to triangulate findings to improve different components of validity.

Qualitative methods, in contrast, are concerned with gaining a deep understanding of phenomenon, which is beneficial for discovering new concepts and studying unique environments and problems (Edmondson & McManus, 2007), features that characterize MTSs operating in extreme environments. In contrast to quantitative research where integrity of findings is dependent on reliability of instrument construction, in qualitative research, “the researcher is the instrument” (Patton, 2002, p. 14). Researchers adopting qualitative methods seek to demonstrate trustworthiness, which is judged against the rules of credibility, transferability, dependability, and confirmability (Lincoln & Guba, 1985).

Credibility refers to the degree to which one can be assured that the researcher’s emergent theory is grounded in the lived experiences of the participants (Polit & Beck, 2012). It is considered to be the most important criterion for establishing trustworthiness and requires the researcher to clearly link study findings to reality. Steps for improving credibility include using multiple qualitative methods, data sources, and observers to check the consistency of findings (triangulation), verifying interpretations with participants, and providing transparent descriptions of experiences, including methods of observation and audit trails (Cope, 2014). Taking these steps goes a long way to also demonstrating transferability, dependability, and confirmability (Lincoln & Guba, 1985). For example, interpretations of initial observations made during live disaster exercises may be compared with other data sources such as interviews with SMEs or with practitioners post-incident or with other researchers who observed the same activities.

Transferability is similar to external validity in terms of being concerned with applying findings to other settings or groups. As the
findings of qualitative research are specific to a small number of participants and environments, demonstrating transferability requires caution to avoid belittling the importance of the contextual factors that impose on the case (Gomm, Hammersley, & Foster, 2000). Providing sufficient thick description of participants, research context, and phenomenon under investigation is important for allowing readers to evaluate whether findings compare to what they see emerging within their situations (Cope, 2014). This includes conveying the boundaries of the study (Shenton, 2004). In live disaster exercises for example, this would include details such as number and location of organizations taking part, restrictions of those who contributed data, number of participants, data collection methods, length and number of data collection sessions, and time period over which data were collected.

Dependability is akin to reliability and refers to the constancy of data across similar conditions (Polit & Beck, 2012). This is achieved by overlapping methods to check for consistency, other researchers concurring with decision trails for each research stage (Cope, 2014), and similar findings being demonstrated with similar participants in similar contexts (Koch, 2006). Study processes therefore need to be reported in detail to enable others to repeat them and evaluate the extent to which appropriate research processes have been followed (Shenton, 2004).

Finally, confirmability refers to the ability of researchers to demonstrate that data represent the responses of participants rather than researcher views and biases (Polit & Beck, 2012). As with dependability, demonstrating confirmability requires researchers to provide an audit trail of processes followed to arrive at interpretations, along with rich participant quotes to show that findings were derived from data (Cope, 2014). Reduced investigator bias is also demonstrated through triangulation, with multiple sources of data or methods verifying findings. For example, by combining detailed observational descriptions of behaviors observed during live disaster exercises with post-incident interviews with practitioners and focus groups or debriefs to compare data from multiple sources.

In summary, live disaster exercises present opportunities to adopt a range of qualitative and quantitative methods to study inter-team processes in situ over time to develop nascent theory of temporal relationships in inter-team processes in MTSs operating in extremis. However, as with other field research contexts, the complexity of live disaster exercises can present challenges for meeting scientific standards, particularly with regard to quantitative methods. The credibility of quantitative methods can be improved by adopting a mixed-method approach to demonstrate how measures relate to key constructs and draw accurate causal conclusions. The credibility of qualitative methods can be improved by adopting a multi-method approach that triangulates methods, data sources, and observers, along with verifying interpretations with participants, transparently describing methods of data collection, participants and context, and providing rich participant quotes to support themes. As will be discussed in the case study section below, there are various ways of implementing these steps to meet the standards appropriate to each methodological class during live disaster exercises.

**Applying principles in practice: A case example**

Disaster response is comprised of leading agencies, such as emergency services, health bodies, and local authorities, and agencies that provide support when a disaster affects their sector such as utility companies (U.K. Civil Contingencies Act, 2004). These MTSs are organized under a three-tiered hierarchical structure where decisions are fed from Strategic Command (responsible for setting overall strategic objectives and agency contributions) to
Tactical (setting operational parameters for utilizing resources available) and finally, Operational (translating tactics into actions to resolve an incident) (Home Office, 2018).

The following case details a data collection framework that was developed for use within a National U.K. Home Office funded exercise referred to as Exercise Joint Endeavour (Ex. JE) to improve the credibility and contribution of findings. The framework was both multi-method (interviews, observations, debriefs) and mixed-method. Multiple qualitative methods were adopted to improve the trustworthiness of findings (see Table 2 for a summary), and mixed-methods were used to strengthen the validity of quantitative methods (Ormerod & Ball, 2010). Data were collected both sequentially (e.g., prior to, during, and post exercise) and concurrently (e.g., collecting multiple sources of data within the exercise), presenting opportunities for knowledge gained from one method to influence and complement another (Creswell, Plano Clark, Gutmann, & Hanson, 2003). For example, analysis of interviews that were conducted with SMEs prior to the exercise taking place as part of the exercise planning phase was used to inform the development of codebooks to count the frequency of key inter-team behaviors and rate key performance criteria.

The framework also adopted both inductive and deductive approaches in line with the state of theory development. Data collection began with an inductive approach, conducting interviews with SMEs to initially identify and define key constructs of importance to the disaster context, which was further supported with a literature review. This was followed by a combination of inductive and deductive approaches to develop knowledge of construct space (the domain it applies to such as the people, events, or objects), conditions relating to onset, duration and cease of processes, and impact on performance. For example, SME interviews informed the development of deductive qualitative measures, such as an interview schedule to examine whether practitioners confirmed findings from SME interviews, and a codebook, and debrief questions. SME interviews also informed the development of deductive quantitative measures such as performance rating scales and frequency counts. Figure 1 provides an overview of the types of data collected prior to, during, and post exercise and how these different forms of data collection informed one another over the course of the study.

**Exercise overview**

Ex. JE was a 9-hr live disaster exercise that involved 1,000 responders from across Police, British Transport Police, Fire, Ambulance, Local Council, National Health Service, Environment Agency, British Red Cross, gas, electricity and water companies, Royal Air Force, and Government. Actors and members of the public played the role of 175 casualties. Five media agencies were also present to generate further realism. The incident ground was a physical reconstruction of a train that had derailed from its tracks, collided with a multistory building (Sector one) and several vehicles and power lines (Sector two), causing a bus to crash into an adult learning center (Sector three). Alongside these physical features, audio, visual, and text-based injects were fed into the exercise to replicate challenges faced during real disasters (e.g., whether to commit crews to a high-risk area or not; what resources to release to assist with a second disaster). Drawing on over 10 years of experience debriefing emergency responders nationally and internationally after numerous real disasters, researchers assisted in developing these realistic challenges to test disaster response policies and procedures.

During Ex. JE, Operational Commanders were based across the three-sector incident ground. Strategic and Tactical Commanders were based at a Command Centre five miles away, as is the usual structure in the U.K. The number of component teams operating on the
**Table 2.** Demonstrating trustworthiness of qualitative methods adopted in studies of live disaster exercises.

| Steps taken                                                                 | How this improves trustworthiness                                                                                   |
|----------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|
| Adopting unobtrusive approach, including naturalistic observations         | Minimizing obtrusiveness of data collection and disruption to exercise delivery improves psychological and physiological immersion, and realism of responses, which is important for linking study findings back to reality to demonstrate that they capture the lived experiences of participants (credibility). |
| Verifying preliminary assumptions with practitioners during hot debriefs and verifying findings and conclusions with practitioners during cold debriefs | Provides an opportunity to confirm the degree to which preliminary assumptions represent participant experiences and perspectives prior to analysis, and interpretations of data post-analysis to improve the trustworthiness of findings (credibility, confirmability). Allows practitioners based in different locations to comment on whether findings apply to the settings they were based in or settings they have previously experienced (transferability). |
| Consultation with experienced practitioners throughout the exercise planning process | Improves ability to identify unobtrusive methods of data collection and opportunities to develop a standardized observational framework that captures features of importance to practice. Facilitates ability to build trust with agencies to gain support and access to practitioners and locations (credibility). |
| Development of standardized observational frameworks                       | Provides consistency in inter-team processes observed across locations, time points and even exercises (credibility). Provides a transparent audit trail for observational data collection that can be adopted by other researchers in similar contexts to verify findings (dependability, transferability). |
| Triangulation of methods and data sources                                  | Integrating methods so that one informs another and provides a more holistic overview of the event and opportunities to collect more detailed and rich accounts to support understanding (credibility, confirmability), for example, observational notes capture aspects that recordings may miss and can be used to improve recollection of interviewees, interviews capture internal cognitive accounts rather than just behaviors and fill in the gaps missed. Allows comparisons to be made across data sources to check for consistencies (credibility, dependability). |
| Triangulation of observers                                                 | Pairing up observers to observe the same settings within an incident and recording incidents so that multiple observers can analyze them post-incident allows comparisons to be made to check for consistencies (credibility, dependability). |
incident ground altered across the course of the incident, in-line with situational demands. For example, the initial operational response consisted of two four-member fire crews, two two-member paramedic crews, and two two-member police teams. The number of teams and agencies involved increased as the scale of the incident was realized, resulting in over 60 component teams of varying sizes. At strategic and tactical levels, the number of component teams remained stable. However, membership across the MTS was fluid due to work shift handovers.

Within this MTS, differentiation was also high with knowledge and capabilities widely distributed across component teams from 13 different public and private sector agencies. There were also diverse specialisms within agencies such as Hazardous Area Response Teams within Ambulance, and Rapid Response and Specialist Operations Response Teams within Fire. Similarly, while agencies shared the superordinate goal of saving lives and reducing risks to public safety, subgoals were diverse. Police sought to preserve evidence for investigative purposes, Fire and Ambulance sought to extract and treat casualties, and businesses sought to reopen as soon as possible. Subgoals also altered across the course of the incident in line with changes in situational awareness. For example, the introduction of a second large incident at a separate site affected goal prioritization and required resources to be split. Teams were faced with the challenge of coordinating and prioritizing the order in which interdependent subgoals were addressed to avoid conflicting actions.

Overall, exercises such as this are beneficial for studying various aspects of temporal relationships in inter-team processes within MTS operating in extremis. This includes identifying the precursors that promote vertical (between different hierarchical levels) and horizontal (between functions within the same hierarchical level) coordination internally and externally within MTSs operating in extreme environments, along with how these relationships alter in line with changes to MTS size and shape, and features of the incident.

**Data collection**

Data were collected prior to, during, and post exercise to study inter-team processes using (i) observations and video recordings (see “During Exercise” section of Figure 1), (ii) semi-structured interviews with SMEs and practitioners responding to the exercise (see “Prior to Exercise” and “During Exercise” sections of...
Data collection overlapped with data from one method informing others, which is beneficial for illuminating relationships that may be missed by a single method (Madill & Gough, 2008). While adopting a range of methods can be time-consuming, it allows phenomenon to be studied from multiple viewpoints to provide robust explanations for complex human behavior. This is beneficial for generating research with real-world applications (Cohen & Manion, 2000).

Although adopting a multi-method and mixed-method approach is not novel in and of itself, it has a novel application in a severely under-researched yet vitally important domain. Live disaster exercises provide valuable opportunities for studying temporal relationships in inter-team processes in extremis. However, a number of important considerations are needed to demonstrate fidelity of data collected within these challenging contexts where complex and dynamic phenomenon is simultaneously co-occurring across a large number of individuals. While considerations discussed here are applicable to all live exercises, disaster exercises are particularly beneficial for studying inter-team processes in extreme environments as they seek to physically and psychologically replicate the complexity of real disasters.

Data collection prior to exercise

Interviews with SMEs. During the exercise-planning phase, interviews were conducted with SMEs from across all three emergency services using an inductive approach. The purpose of these interviews was to develop a deeper understanding of (i) the inter-team processes of importance for improving coordination during disaster response and (ii) key indicators of effective performance within these contexts (phase one—see “Prior to Exercise” section of Figure 1). Interviews were transcribed and analyzed prior to the exercise using qualitative thematic analysis, and the findings of this analysis informed the development of several other qualitative and quantitative data collection methods that were used during and post exercise, as detailed below.

Data collection during exercise

Observational methods. The use of observational methods is beneficial for exploring demands and strategies that people have developed to cope in particular contexts (Crandall et al., 2006). Although rare, observational research into inter-team processes in MTSs during live disaster exercises provides a useful overview of obstacles, challenges (Bharosa et al., 2010), and how technologies can be used to coordinate activities (Gonzalez, 2008). However, such research has been less transparent regarding how data were collected and analyzed to arrive at empirical claims. This poses implications for evaluating whether they truly represent participant views and experiences rather than researchers’ views and biases. Lack of detail regarding how data were collected also affects ability to adopt similar approaches in similar contexts. Providing a transparent account of steps taken to collect data is also important for making comparisons across studies to develop MTS theory regarding what factors affect the onset, duration, and changes in phenomena and why.

During Ex. JE, a number of steps were taken to improve the trustworthiness of observational data collected. All methods were selected based on their ability to minimize obtrusiveness. Exercise realism was important for psychologically and physically immersing practitioners so that they responded as they would in a real disaster. For example, multiagency meetings held to coordinate information, risk assessments, strategies, and actions were recorded using static cameras at Strategic and Tactical levels, and Commanders wore helmet-mounted cameras at Operational levels. While limited battery life currently raises the dilemma of whether to focus on recording the first few
hours of Operational Command or risk disrupting the exercise to replace batteries, technological advances in battery life and discreteness of equipment are improving.

Post-exercise, all recordings were transcribed and time stamped, providing a record of inter-team processes, when and how they manifest and altered. These data have been coded both qualitatively (inductive phase one—descriptions of behaviors observed across the course of the exercise to identify and define key constructs) and quantitatively (deductive phase two—counting frequency of occurrence for key defined constructs). Multiple researchers were able to code recordings at a later date to improve trustworthiness through triangulation of observers. However, it is important to note that while recordings are beneficial for examining behavior, they do not indicate underlying cognitions. To address this, recordings can be shown to participants post-exercise to improve the accuracy and depth of their internal reflections of cognitions and affects (Cohen-Hatton, Butler, & Honey, 2015). Improving the depth and accuracy of participant accounts is important for demonstrating credibility and providing rich participant quotes to support themes.

Recordings are also limited in that they only capture what is happening in the direction cameras are pointed at, which can result in a misleading or cognitively narrowing account (Crandall et al., 2006). For example, static cameras provided an objective “live account” of aspects of the incident but did not capture smaller ad hoc meetings or radio communications. Accordingly, a team of 16 observers also kept observational notes to provide wider context during the complex and dynamic event simultaneously co-occurring across a large number of individuals and locations. To improve consistency in the inter-team processes observed across locations over time, a standardized coding framework was designed prior to Ex. JE (deductive phase two), primarily based on SME interviews, but also compared against a detailed literature review to ensure academic and practical relevance.

The standardized coding framework involved both a set of key inter-team behaviors and a set of exercise performance indicators. Observers were required to score inter-team behaviors on a scale of 0 (completely absent) to 2 (consistently present) and provide qualitative descriptions of activities observed relating to these behaviors as context. Observers were required to score performance indicators on a scale of 1 (very poor) to 7 (very good) and provide a description of why this score had been given to provide context. Observers completed coding frameworks a minimum of once an hour to capture changes in behaviors over time. While it would be beneficial to capture observations in situ more frequently, keeping these records is cognitively demanding and runs the risk of missing observing key behaviors. Accordingly, coding frameworks were supplemented with the video recordings, which were coded using an inductive thematic approach post-exercise to make finer discriminations.

This same framework has now been applied across multiple live disaster exercises. Difficulties in making comparisons across exercises pose implications for developing theory and assessing whether interventions implemented to improve inter-team practices are effective outside of classroom settings within the complex environments emergency responders operate. Adopting standardized frameworks across exercises is beneficial in allowing comparisons to be made (e.g., between the number, size and diversity of component teams, the type of incident, or impact of interventions). It also provides a means of collecting quantitative data of complex interactions such as frequency of observed behaviors and ratings of performance effectiveness (Crandall et al., 2006). However, there is a danger of discarding significant data if it does not relate to a proscribed checklist and observers are unaware of its significance, making it important to combine this method with other data such as recordings and interviews.
Ensuring that all observers are skilled in identifying and recording inter-team processes prior to data collection is also important for demonstrating trustworthiness. In their study of shared leadership in six-member MTS aircrews during flight simulations, Bienefeld and Grote (2013) showed observers video footage of interactions in contexts similar to the ones they would collect data and compared their coding using interrater reliability to assess effectiveness of preparatory steps. For researchers involved in multiple live exercises, a similar approach can be adopted by drawing on recordings from previous exercises as a source of observational training materials. For Ex. JE, in addition to providing observational training, pairs of observers coded the same interactions during the exercise. Half were academic researchers specializing in team processes in risky and uncertain contexts and half were emergency responders with between eight and 37 years of practical experience. Academics and practitioners were paired to provide independent observations of the same interactions. Observations were compared for concurrence post-incident to improve trustworthiness (Armstrong, Gosling, Weinman, & Marteau, 1997).

In addition, the 16 observers used a secure instant group messaging system to share details of activities observed across locations. The focus here was on providing descriptions of activities rather than interpretations of performance as this could influence how others viewed the exercise. Examples of the types of text-based messages shared include “The first two fire service appliances have arrived at sector two containing eight firefighters” and “Ambulance tactical commander is ending his shift.” Text, audio, photographic, and video messages were shared securely as information was encrypted end-to-end, and messages were automatically time stamped. Sharing this information across locations served a dual purpose. Firstly, it enabled researchers to build a global picture of the incident to coordinate data collection activities, such as conducting interviews with practitioners when there was a shift change over. Information could also be used to inform interviews by flagging significant activities (e.g., “A media briefing has been called for 2pm”). Secondly, this time stamped record of activities could be used post-incident to support researchers in producing detailed descriptions of incident context for publications.

**Interviews with practitioners.** Interviews provide a flexible data collection approach for gathering detailed and rich information, especially if the interviewer is sensitive to contextual variations in meaning (Phellas, Bloch, & Seale, 2011). Obtaining accounts after an event can provide useful insights into cognitive causes of difficulty and success, as well as practical learning and problem-solving benefits (VanLehn, Jones, & Chi, 1991). Interviews can also be used to address logistical complications such as failing to conduct an observation at the right moment, meaning key dynamics and elements of task performance are missed (Crandall et al., 2006).

One method commonly used for gaining deeper understanding of underlying procedural knowledge is think-aloud protocols (Ormerod & Ball, 2010). However, such methods are not practical for group activities or live exercises because they affect the realism of behaviors and interactions, and the processing effort required can distort task performance (Ericsson & Simon, 1993). An alternative approach that has been used within live disaster exercises is on-the-spot interviews, which assist with gaining better understanding of responders’ actions and interpretations of problems (Bharosa et al., 2010). However, this method was not utilized in Ex. JE because it also runs the risk of distracting attention from completing tasks. Instead, interviews were scheduled to take place during shift changeovers and immediately post-exercise to capture initial reflections.

*Interview schedules (deductive phase two)* were designed by drawing on analysis of SME interviews and literature review to inform the
focus of the questions asked. This allowed comparisons to be made between practitioners’ experiences and perspectives of important inter-team processes, performance indicators, and the relationships between these components and those of SMEs and the wider literature. Interviews were also shaped by observations conducted during the exercise, which allowed researchers to tailor additional questions to gain valuable direct internal perceptions of inter-team processes that improved and hindered the disaster response within the live exercise and why (Alvesson, 2011).

**Data collection post-exercise**

*Debriefs.* Also known as “after action reviews,” debriefs were designed by the U.S. military to improve learning and performance (Tannenbaum & Cerasoli, 2013). The aim is to lead a group of practitioners through a series of questions to discuss actions and thought processes in a nonpunitive environment that encourages reflection on recent experience (Kessler, Cheng, & Mullan, 2015). Debriefs allow multiple perspectives to be gathered at a time, and comments made by one person may aid the memory of another. They can also enhance performance by assimilating improved behaviors in practice (Tannenbaum & Cerasoli, 2013).

Debriefs typically occur in two forms. “Hot” debriefs are held immediately following the event to capture immediate thoughts and feelings (Kessler et al., 2015). “Cold” debriefs are held sometime after to capture more in-depth perspectives once participants have had the opportunity to process what occurred, providing contrasting insights (Tannenbaum & Cerasoli, 2013). Both are valuable for encouraging people to construct their own meanings for their actions and to aid in identifying lessons (Allen, Reiter-Palmon, Crowe, & Stott, 2018), making them valuable for gathering perspectives of individual, team, and inter-team responses. They are also useful for improving the credibility of research with hot debriefs providing a means of verifying researchers initial impressions and cold debriefs a means of verifying conclusions drawn after in-depth analysis.

For Ex. JE, three “hot” debriefs were conducted immediately post-exercise with practitioners from across agencies involved with the response. These debriefs were led by exercise planners but questions were shaped by the research team based on initial observations of the exercise and comments raised during interviews. One month post-exercise, “cold” debriefs were conducted with 150 practitioners from across agencies who were organized into smaller groups of 10 to capture perspectives on the way they had collaborated to resolve the incident (deductive phase two). These debriefs were led by the research team using a semi-structured schedule developed based on post-incident analysis of observational data (both standardized coding framework and video recording) and interview data. The purpose of taking this deductive approach in tailoring questions was to verify conclusions and clarify points.

**Data analysis and integration**

The multiphase nature of data collection prior to, during, and post exercise allowed data integration to take place at the data collection, data analysis, and interpretation stages of inquiry (see Waring et al., 2018 for an account of how data were integrated, analyzed, and conclusions drawn from findings). This allowed the research approach to progress from inductive to deductive, presenting opportunities to both define key constructs, including their space, onset, alteration, and cease of processes, and to provide preliminary tests of this.

For example, standardized coding frameworks provided a tool for integrating findings within the data collection phase. Scores for frequency of behaviors and performance ratings were compared against the qualitative
descriptions provided as justification for these scores to highlight whether there were any particular patterns or precursors for the occurrence of such behaviors. At the data analysis stage, video recordings were transformed into numerical frequencies of behavior and performance ratings. This presented an opportunity to test whether there were patterns between the frequency of particular behaviors and performance indicators and compare whether these same relationships were present in both the data collected using the standardized coding framework and the video footage. The inductive thematic analysis of video recordings, interviews, and debriefs was integrated with the standardized coding framework at the data interpretation phase to verify whether this framework captured all of the important constructs or whether there were any additional processes and performance indicators that were important to consider.

Taken together, data collected using these combinations of methods and approaches provide a richer understanding of inter-team processes that are important for effective MTS performance in extreme environments, the contextual features that lead to the onset and cease of these processes and how this impacts performance. Such knowledge can be used to inform hypothesis generation and testing, providing new avenues for future research. It also allows some preliminary conclusions to be drawn regarding temporal frameworks for informing how frequently future measures should be implemented to capture the presence of phenomena. Similarly, observational recordings can be coded and analyzed using lag sequential analysis to analyze patterns of interactions between inter-team behaviors. Finally, this data collection approach has resulted in the development of a standardized coding framework that can be used across live disaster exercises to make comparisons and further refine understanding of the boundaries under which inter-team processes improve and hinder MTS functioning. Standardized coding frameworks such as this also serve as a valuable tool for practitioners in evaluating their performance, identifying specific areas where development is required, and testing the impact of interventions on performance in challenging contexts.

**Conclusion**

Within large MTSs operating in disasters and other extreme environments, coordination has repeatedly been highlighted as problematic, compromising the effectiveness of response. To date, inter-team processes have largely been studied as static factors due to a lack of longitudinal research to identify growth trajectories and fluctuations over time. As a result, temporal theories of inter-team processes in MTSs remain nascent. This poses implications for understanding and addressing the problem of coordination difficulties in MTS operating in extreme environments. Research is needed that focuses on MTSs in the contexts they exist, and studies change in inter-team processes over time to advance nascent theories and develop evidence-driven targeted interventions to improve MTS functioning in extreme environments (Shuffler & Carter, 2018; Waring et al., 2018; Waring, Alison, Humann, & Shortland, 2019). In contrast to both laboratory-based studies and real disasters, live disaster exercises offer the potential to study inter-team processes in situ in extremis.

However, using live disaster exercises to conduct research to develop nascent theory is not without its challenges in terms of ensuring the credibility of findings. Accordingly, this article presented a data collection framework that is both multi-method and mixed-method and adopts inductive and deductive approaches to promote methodological (Edmondson & McManus, 2007) and measurement fit (Luciano et al., 2018). The article also discussed considerations for improving the credibility of data collected, including working alongside practitioners throughout the exercise.
planning process to identify suitable unobtrusive measures and develop standardized coding frameworks that can be applied across multiple exercises to assist with testing interventions and strengthening MTS theory by making comparisons. There is still a need to continuously develop, revise, and improve methods to adapt to the increasingly complex problems people face. But the abundance of data collection opportunities in live disaster exercises can support future development of complex models to improve performance in extreme events that threaten safety and security.

Nonetheless, a key limitation to be aware of in using live disaster exercises to study inter-team processes is the level of resource required. This includes the potential amount of time researchers may need to spend embedding themselves within agencies and planning meetings to develop trust and understanding of the practices of the organization(s) in relation to the exercise and the way it will run. It also includes the level of resources and number of researchers that may be needed to collect data using a wide variety of methods across the course of an exercise. Researchers may need to be flexible in adopting a range of different methods depending on the dynamics of the exercise, which makes it vital to ensure that all have received appropriate training and have knowledge regarding the study context. Similarly, analyzing and making sense of this complex data post-exercise can be a time- and resource-intensive process, including transcribing audio and visual footage and adopting qualitative and quantitative methods of analysis to examine constructs and present findings to address the research question.

It is also important to note that live exercises are expensive for the agencies that are designing and delivering them to test their policies, procedures, and responses to unique events. Evaluating performance within these complex and dynamic environments can be difficult to achieve, which limits the ability of agencies to systematically identify whether improvements have been made and where further focus is required. Development of standardized coding frameworks such as the one developed through this study can serve as a valuable resource for agencies to apply across all live exercises to provide consistency and allow direct comparisons to be made, which is important for improving practice.

Acknowledgement

The author would like to thank the two anonymous reviewers, the special issue editors, Matthew Cronin and Marissa Shuffler, and Marie Eyre for their time and support in shaping this article. Their advice throughout the review process was invaluable for substantially strengthening the focus of the article.

Funding

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

ORCID iD

Sara Waring https://orcid.org/0000-0003-1625-5705

References

Abernethy, M. A., Chua, W. F., Luckett, P.F., & Selto, F. H. (1999). Research in managerial accounting: Learning from others’ experiences. Accounting & Finance, 39, 1–27.

Allen, J. A., Reiter-Palmon, R., Crowe, J., & Stott, C. (2018). Debriefs: Teams learning from doing in context. American Psychologist, 73, 504–516.

Alvesson, M. (2011). Interpreting interviews. Sage.

Berlin, J. M., & Carlström, E. D. (2015). Collaboration exercises: What do they contribute? A study of learning and usefulness. Journal of Contingencies & Crisis Management, 23, 11–23.

Bharosa, N., Lee, J., & Janssen, M. (2010). Challenges and obstacles in sharing and coordinating information during multi-agency disaster
response: Propositions from field exercises. *Information Systems Front*, 12, 49–65.

Bienefeld, N., & Grote, G. (2013). Shared leadership in multiteam systems: How cockpit and cabin crews lead each other to safety. *Human Factors*, 56, 270–286.

Black, T. R. (1999). *Doing quantitative research in the social sciences*. Sage.

Brehmer, B., & Dorner, D. (1993). Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study. *Computers in Human Behavior*, 9, 171–184.

Campbell, J. P. (Ed) (1990). *The role of theory in industrial/organizational psychology* (Vol. 2). Consulting Psychologists Press.

Carter, D. R. (2014). The impact of leadership network structure on multiteam system innovation. Unpublished thesis, Georgia Institute of Technology, Atlanta.

Casey, D., & Murphy, K. (2009). Issues in using methodological triangulation in research. *Nurse Researcher*, 16, 40–55.

Cialdini, R. B. (1980). Full-cycle social psychology. *Applied Social Psychology Annual*, 1, 21–47.

Civil Contingencies Act. (2004). *Local arrangements for civil protection*. Retrieved from https://www.legislation.gov.uk/ukpga/2004/36/contents

Cobb, M. (1999). The impact of environmental complexity and team training on team processes and performance in multiteam systems. Unpublished doctoral dissertation, The Pennsylvania State University, University Park, PA.

Cohen, L., & Manion, L. (2000). *Research methods in education* (5th ed.). Routledge.

Cohen-Hatton, S. R., Butler, P. C., & Honey, R. C. (2015). An investigation of operational decision making in situ: Incident command in the UK fire and rescue service. *Human Factors*, 57, 793–804.

Cope, D. G. (2014). Methods and meanings: Credibility and trustworthiness of qualitative research. *Oncology Nursing Forum*, 41, 89–91.

Corley, K. G., & Gioia, D. A. (2011). Building theory about theory building: What constitutes a theoretical contribution? *Academy of Management Review*, 36, 12–32.

Crandall, B., Klein, G., & Hoffman, R. R. (2006). *Working minds: A practitioner’s guide to cognitive task analysis*. MIT Press.

Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 209–240). Sage.

Cronin, M. A., Weingart, L. R., & Todorova, G. (2011). Dynamics in groups: Are we there yet? *The Academy of Management Annals*, 5, 571–612.

Crowe, J., Allen, J. A., & Bowes, B. (2014). Multicrew responses to a structure fire: Challenges of multiteam systems in a tragic fire response context. In M. L. Shuffler, E. Salas, & R. Rico (Eds.), *Pushing the boundaries: Multiteam systems in research and practice* (pp. 205–219). Emerald.

Davison, R. B., Hollenbeck, J. R., Barnes, C. M., Sleesman, D. J., & Ilgen, D. R. (2012). Coordinated action in multiteam systems. *Journal of Applied Psychology*, 97, 808–824.

de Vreede, G. J., Briggs, R. O., & Reiter-Palmon, R. (2010). Exploring asynchronous brainstorming in large groups: A field comparison of serial and parallel subgroups. *Human Factors: The Journal of the Human Factors & Ergonomics Society*, 52, 173–188.

de Vries, T. A., Hollenbeck, J. R., Davison, R. B., Walter, F., & Van der Vegt, G. S. (2016). Managing coordination in multiteam systems: Integrating micro and macro perspectives. *Academy of Management Journal*, 59, 1823–1844.
multiteam environments. *The Leadership Quarterly*, 22, 152–169.

DeConstanza, A., DiRosa, G., Jiménez-Rodríguez, M., & Cianciolo, A. (2014). No mission too difficult: Army units within exponentially complex multiteam systems. In M. L. Shuffler, E. Salas, & R. Rico (Eds.), *Pushing the boundaries: Multiteam systems in research and practice* (pp. 61–76). Emerald.

DiRosa, G. (2013). *Emergent phenomena in multiteam systems: An examination of between-team cohesion*. Unpublished doctoral dissertation, George Mason University.

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research & Perspectives*, 38, 105–124.

Edmondson, A. C., & McManus, S. E. (2007). Methodological fit in management field research. *The Academy of Management Review*, 32, 1155–1179.

Eisenhardt, K., Graebner, M., & Sonenshein, S. (2016). Grand challenges and inductive methods: Rigor without rigor mortis. *Academy of Management Journal*, 59, 1113–1123.

Ericsson, K., & Simon, H. A. (1993). Verbal reports as data. *Psychological Review*, 87, 215–251.

Fine, G. A., & Elsbach, K. D. (2000). Ethnography and experiment in social psychology theory building: Tactics for integrating qualitative field data with quantitative lab data. *Journal of Experimental & Social Psychology*, 36, 51–78.

Firth, B. M., Hollenbeck, J. R., Miles, J. E., Ilgen, D. R., & Barnes, C. M. (2015). Same page, different books: Extending representational gaps theory to enhance performance in multiteam systems. *Academy of Management Journal*, 58, 813–835.

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The Qualitative Report*, 8, 597–607.

Goldthorpe, J. H. (2000). *On sociology: Numbers, narratives, and the integration of research and theory*. Oxford University Press.

Gomm, M., Hammersley, M., & Foster, P. (2000). *Case study method*. Sage.

Gonzalez, R. A. (2008). Coordination and its ICT support in crisis response: Confronting the information-processing view of coordination with a case study. *Proceedings of the 41st Hawaii International Conference on System Sciences*, 1–10.

Hambrick, D. C. (2007). The field of management’s devoting to theory: Too much of a good thing? *Academy of Management Journal*, 50, 1346–1352.

Healey, M. P., Hodgkinson, G. P., & Teo, S. (2009). Responding effectively to civil emergencies: The role of transactive memory in the performance of multiteam systems. *Proceedings of NDM9, the 9th International Conference on Naturalistic Decision Making*.

Home Office. (2018). *Critical incident management (version 12)*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/736743/critical-incident-management-v12.0ext.pdf.

Issenberg, S. B., & Scalese, R. J. (2008). Simulation in health care education. *Perspectives in Biology & Medicine*, 51, 31–46.

Joppe, M. (2000). *The research process*. Retrieved from http://www.ryerson.ca/~mjoppe/rp.htm

Kerslake, R. W. (2018). *The Kerslake Report: An independent review into the preparedness for, and emergency response to, the Manchester Arena attack on 22nd May 2017*. Retrieved March 7, 2019, from https://www.kerslakearenaview.co.uk/media/1022/kerslake_arena_review_printed_final.pdf

Kessler, D. O., Cheng, A., & Mullan, P. C. (2015). Debriefing in the emergency department after clinical events: A practical guide. *Annals of Emergency Medicine*, 65, 690–698.

Koch, T. (2006). Establishing rigour in qualitative research: The decision trail. *Journal of Advanced Nursing*, 53, 91–100.

Koch, T., & Harrington, A. (1998). Reconceptualising rigour: The case for reflexivity. *Journal of Advanced Nursing*, 28, 882–890.

Kozlowski, S. W. J. (2015). Advancing research on team process dynamics: Theoretical, methodological, and measurement considerations. *Organizational Psychology Review*, 5, 270–299.
Lanaj, K., Hollenbeck, J. R., Ilgen, D. R., Barnes, C. M., & Harmon, S. J. (2013). The double-edged sword of decentralized planning in multiteam systems. *Academy of Management Journal*, 56, 735–757.

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

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage.

Lipshitz, R. (2010). Rigor and relevance in NDM: How to study decision making rigorously with small Ns and without controls and (inferential) statistics. *Journal of Cognitive Engineering & Decision Making*, 99–112.

Luciano, M. M., DeChurch, L. A., & Mathieu, J. E. (2015). Multiteam systems: A structural framework and meso-theory of system functioning. *Journal of Management*, 44, 1065–1096.

Luciano, M. M., Mathieu, J. E., Park, S., & Tannenbaum, S. I. (2018). A fitting approach to construct measurement alignment: The role of big data in advancing dynamic theories. *Organizational Research Methods*, 21, 592–632.

Lyons, E., & Coye, A. (2016). *Analysing qualitative data in psychology*. Sage.

Madill, A., & Gough, B. (2008). Qualitative research and its place in psychological science. *Psychological Methods*, 13, 254–271.

Majchrzak, A., Jarvenpaa, S. L., & Hollingshead, A. B. (2007). Coordinating expertise among emergent groups responding to disasters. *Organization Science*, 18, 147–161.

Marks, M. A., DeChurch, L. A., Mathieu, J. E., Panzer, F. J., & Alonso, A. (2005). Teamwork in multiteam systems. *Journal of Applied Psychology*, 90, 964–971.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, 26, 356–376.

Marvasti, A. B. (2014). *The SAGE handbook of qualitative data analysis*. Sage.

Mathieu, J. E. (2016). The problem with [in] management theory. *Journal of Organizational Behavior*, 37, 1132–1141.

McMillan, J. H. (2007). Randomized field trials and internal validity: Not so fast my friend. *Practical Assessment, Research & Evaluation*, 12. Retrieved from http://pareonline.net/getvn.asp?v=12&n=15

Meyer, S. B., & Lunnay, B. (2013). The application of abductive and retroductive inference for the design and analysis of theory driven sociological research. *Sociological Research Online*, 18, 12.

O’Sullivan, A. (2003). Dispersed collaboration in a multi-firm, multi-team product development project. *Journal of Engineering & Technology Management*, 20, 93–116.

Ormerod, T., & Ball, L. (2010). Cognitive psychology. In C. Willig & W. Stainton-Rogers (Eds.), *The SAGE handbook of qualitative research in psychology* (pp. 554–576). Sage.

Patrick, J. (2011). Haiti earthquake response: Emerging evaluation lessons. *Evaluation insights*. Network on Development Evaluation of the OECD Development Assistance Committee.

Patton, M. Q. (2002). *Qualitative evaluation and research methods* (3rd ed.). Sage.

Phellas, C. N., Bloch, A., & Seale, C. (2011). Structured methods: Interviews questionnaires and observation. In C. Seale (Ed.), *Researching society and culture* (pp. 181–205). London: Sage Publications.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2016). Recommendations for creating better concept definitions in the organizational, behavioral, and social sciences. *Organizational Research Methods*, 19, 159–203.

Polit, D. F., & Beck, C. T. (2012). *Nursing research: Generating and assessing evidence for nursing practice*. Lippincott Williams and Wilkins.

Pollock, K. (2013). Review of persistent lessons identified relating to interoperability from emergencies and major incidents since 1986, emergency planning college occasional papers, number 6. Retrieved from http://www.jesip.org.uk/wp-content/uploads/2013/07/Pollock-Review-Oct-2013.pdf
Popper, K. (2002). *The logic of scientific discovery*. Routledge.

Schwarz, G., & Stensaker, I. (2014). Time to take off the theoretical straightjacket and (re-)introduce phenomenon-driven research. *The Journal of Applied Behavioral Science*, 50, 478–501.

Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. *Education for Information*, 22, 63–75.

Shuffler, M. L., Jiménez-Rodríguez, M., & Kramer, W. (2015). The science of multiteam systems: A review and future research agenda. *Small Group Research*, 46, 659–699.

Shuffler, M. L., & Carter, D. R. (2018). Teamwork situated in multiteam systems: Key lessons learned and future opportunities. *American Psychologist*, 73, 390–406.

Smith, W., Dowell, J., & Ortega-Lafuente, M. A. (1999). Designing paper disasters: An authoring environment for developing training exercises in integrated emergency management. *Cognition, Technology & Work*, 1, 19–131.

Sutton, R. I., & Staw, B. M. (1995). What theory is not. *Administrative Science Quarterly*, 40, 371–384.

Tannenbaum, S. I., & Cerasoli, C. P. (2013). Do team and individual debriefs enhance performance? A meta-analysis. *Human Factors*, 55, 231–245.

Thorne, S. (2000). Data analysis in qualitative research. *Evidence-Based Nursing*, 3, 68–70.

Trochim, W. M. K. (2006). Introduction to validity. *Social Research Methods*. Retrieved April 11, 2019, from www.socialresearchmethods.net/kb/introval.php

VanLehn, K., Jones, R. M., & Chi, M. (1991). A model of the self-explanation effect. *Journal of the Learning Sciences*, 1, 69–106.

Waring, S., Alison, L., Humann, M., & Shortland, N. (2019). The role of information sharing on decision delay during multiteam disaster response. *Cognition, Technology & Work*. Advance online publication. doi:10.1007/s10111-019-00570-7

Waring, S., Alison, L., McGuire, G., Barrett-Pink, C., Humann, M., Swan, L., & Zilinsky, T., (2018). Information sharing in inter-team responses to disaster. *Journal of Occupational & Organizational Psychology*, 91, 591–619.

Weick, K. E. (1979). *The social psychology of organizing*. McGraw-Hill.

Yauch, C. A., & Steudel, H. J. (2003). Complementary use of qualitative and quantitative cultural assessment methods. *Organizational Research Methods*, 6, 465–481.

**Author biography**

**Sara Waring** is a lecturer in Investigative and Forensic Psychology, and Research Coordinator of the Critical and Major Incident research group at the University of Liverpool. She received her PhD from the University of Liverpool in 2012. Sara has worked alongside law enforcement, emergency services, and associated agencies for over a decade, studying decision-making, communication, and information sharing in risky and uncertain contexts, and feeding findings back into practice through the design, delivery and evaluation of training, policy and guidance.