Immersive simulations with Extreme Teams

Olivia Brown*, Nicola Power, & Stacey M. Conchie

*Correspondence should be directed to oab42@bath.ac.uk

Abstract

Extreme teams work in challenging, high pressured contexts, where poor performance can have severe consequences. These teams must coordinate their skill sets, align their goals, and develop shared awareness; all under stressful conditions. How best to research these teams poses unique challenges as researchers seek to provide applied recommendations whilst conducting rigorous research to test how teamwork models work in practice. In this paper we identify immersive simulations as one solution to this, outlining their advantages over existing methodologies and suggesting how researchers can best make use of recent advances in technology and analytical techniques when designing simulation studies. We conclude that immersive simulations are key to ensuring ecological validity and empirically reliable research with extreme teams.

Keywords: *Teamwork, Simulations, Extreme teams*

Immersive Simulations with Extreme Teams

'Extreme teams' (ETs) operate in challenging environments in which there are considerable physical, psychological and interpersonal demands (Manzey & Lorenz, 1999). ETs share many similarities with 'High Reliability Organisations', in which teams are required to operate effectively, in complex task environments, and for sustained periods of time (Roberts, 1990; Klein, Bigley & Roberts, 1995). What both contexts have in common, and what defines an ET, is that they operate in atypical environments (in terms of demands/stress levels), in which ineffective performance can have severe, potentially life or death, consequences (Bell, Fisher, Brown & Mann, 2018). Examples of ETs include those involved in long-duration space flights (Zhang et al., 2018), submarine command and control rooms (Stanton & Roberts, 2018), medical emergencies (Klein et al., 2006), high-risk industries (Sneddon, Mearns & Flin, 2006) and emergency response (Power & Alison, 2017a). Interest in ETs is increasing (see Driskell, Driskell & Salas, 2018; Roma & Bedwell, 2017), with teamwork viewed as a vital component to organisational success and safe working practices (Hughes et al., 2016; Mazzocco et al., 2009). This has led to a consideration of how to study these unique, often hard to reach teams, to conduct rigorous applied research that contributes to wider theoretical understanding (Bell et al., 2018; Kozlowski, 2015). Given the unique context in which ETs operate, this understanding may diverge from what we know about conventional teams and challenge our current thinking. We identify immersive simulations as one way to achieve this and present a framework for designing, conducting and analysing this research, drawing on current research and ethnographic experience.

Researching extreme teams

Teamwork is essential for safety and success in extreme environments (Hughes et al., 2016). For example, research in high-risk industries has shown that accidents occur more often due to problems between team members than unsafe working conditions (Dwyer & Raftery, 1991); a finding that has been attributed to issues around poor leadership (McCabe, Loughlin, Munteanu, Tucker & Lam, 2008) and a lack of team spirit (Kadiri et al., 2014). Risser, Rise, Salisbury, Simon and Berns (1999) also showed from fifty-four incidents across eight US hospital emergency departments that half of all recorded deaths and permanent disabilities could have been prevented through better teamwork. Identifying solutions to improve teamwork in ETs can be challenging. This is because they have complex team structures, often form (and dismantle) rapidly, draw on multiple agencies and operate in dynamic conditions that impose a high level of stress on members due to the severe consequences of poor teamwork (Crichton, Flin & Rattray, 2000; Schmutz, Lei, Eppich & Manser, 2018). These features are different to what we see in conventional teams and suggest that theoretically, their processes may be structured differently.

Research on teams requires careful consideration of the complex interplay between performance and its antecedent factors that reside at four levels: the individual (e.g., personality), the team (e.g., team structure: horizontal or vertical), cultural (e.g., organisational culture) and contextual (e.g., task demands). Each of these levels, in isolation and in combination, influence how well a team adapts and responds to a situation. When applied to ETs, an extra layer of complexity is added when we consider the extent to which psychological pressures (e.g., stress) interact with each of these levels and alters team performance (Driskell et al., 2018). The experience of stress can create a perception that task demands exceed available resources, which can lead to undesirable physiological, psychological, behavioural and/or social outcomes (Salas, Driskell & Hughes, 1996). These demands may reside in conventional teams to a lesser extent (or not at all), or in a qualitatively different way (e.g., relating to performance rather than the loss of life). Differences in contextual demands can drive the *type* of stress experienced in teams, which may change or amplify the drivers of effective teamwork (Maynard, Kennedy & Resick, 2018; Driskell et al., 2018). Considering this, researchers have called for empirical research with ETs to test if theoretical models developed with conventional work teams apply to those working in these challenging settings (Vessey & Landon, 2017), and to develop solutions that can protect workers and enhance performance (Power, 2018).

Simulation research with extreme teams

Researchers looking at ETs have employed a variety of methods to understand their composition, function and processes. When the research question concerns a descriptive understanding of ETs, qualitative methods such as observations and interviews (used in isolation or together), have been shown to be effective. Gillespie, Gwinner, Chaboyer and Fairweather (2013), for example, developed an ethnographic account of surgical teamwork culture using observations and interviews. Power and Alison (2017a) identified nine core challenges for commanders during emergencies using interviews. When the research question concerns the influence of self-perceptions on teamwork, self-report measures such as questionnaires have been used. Wauben et al. (2011) found differences between medical team members' in the way they perceived non-technical skills (e.g., communication and situation awareness) using a questionnaire survey. However, what these studies do not do, and what is distinct in simulation studies, is manipulate specific variables to test theory and generate empirical evidence of how these variables influence team performance. Whilst the manipulation of variables is possible in traditional laboratory studies, these studies often utilise student samples in a setting that is void of the stressors present in an extreme environment (e.g., Zaccaro et al., 1995). Further, research highlights the importance of expertise in extreme environments (Boulton & Cole, 2016), thus suggesting that understanding how practitioners work in the real-world necessitates that research is undertaken with the population of interest.

One effective method for studying ETs are simulations. Simulations allow for the measurement of complex relationships between factors that impact team performance in a meaningful organisational context, whilst facilitating a high level of experimental control (Alison et al., 2013; Manser, Dieckmann, Wehner & Rall, 2007). Example relationships may include the impact on performance of individual differences (e.g., attitudes), trust between team members, temporal patterns in teamwork over time, and cultural and contextual variables that may moderate these relationships, such as organisational norms and task demands. Studies that have used simulations to answer such questions include Bienefeld and Grote (2014) who showed the influence of expertise and organisational knowledge on leadership behaviours in aviation teams; and Amacher et al. (2017) who demonstrated that all-female medical teams showed less “hands-on” time and a greater delay before chest compressions in comparison to all-male teams.

In comparison to alternative methods, simulations have five key benefits; they: (i) re-create the stressors and challenges of the workplace; (ii) involve data collection with the population of interest (i.e., practitioners instead of students); (iii) provide an opportunity for researchers to test theory by manipulating and measuring discrete variables; (iv) allow for the collection of rich quantitative and qualitative data related to team behaviour in real time, and (v) can be used as a training tool to increase participation (Rosen et al., 2008). Simulations are an especially useful platform for collecting data with ETs as they provide a physiologically and psychologically safe space that will not endanger participants (Alison et al., 2013), whilst eliciting similar behavioural patterns as would be found *in situ* (Manser et al., 2007). They are also suited to research with ETs who may be difficult to study using alternative methods (e.g., the security sensitive nature of military command control would negate an observation study).

This paper has two main aims. Firstly, it seeks to show the utility of immersive simulations in studying a range of ETs; not just those who operate in healthcare, where many of the frameworks and benefits of utilising immersive simulations originate (see Cheng et al., 2016, Cheng et al., 2014). We will show in this paper that they can also be in contexts where ETs are less well-structured (e.g., multi-team systems), more fluid (e.g., non-stable team members) and involve both horizontal (i.e., within an operational team) and vertical (i.e., between operational, tactical and strategic teams) organisational structures. Secondly, the paper will outline recent technological and analytical advances in psychological research and consider how simulation research can be improved by utilising more immersive methods that can better harness these advances. For example, by considering in what way emerging virtual reality technologies or alternative statistical approaches (i.e., Bayesian statistics) might be used to allow advanced models of ETs to develop. These developments have implications beyond the ET context and hold promise for team research in general. In this paper, we address these aims by outlining a framework for using immersive simulations for research with ETs, broadly focussing on three aspects of the research lifecycle: (i) simulation design, (ii) data collection; and (iii) data analyses.

Simulation design

A simulation seeks to create a testing environment that closely replicates reality (Sleeper & Thomspon, 2008). An important consideration during research design is how to embed fidelity and immersion so that participants feel engaged in the simulation and exhibit similar behaviours as would be found *in situ*. Fidelity and immersion are two inter-related constructs that seek to increase the sense of realism during a simulation (Alison et al., 2013; Lester, Georgiou, Hein, Littlepage, Moffet III & Craig, 2017), and which determine the success of simulations. *Fidelity* is the extent to which the simulation matches the real-world environment (Maran & Glavin 2003). This influences the level of *immersion* felt by the participant, defined as the "subjective impression that one is participating in a comprehensive, realistic experience" (Dede, 2009, p. 66). Fidelity can be created at the physical and psychological levels. Physical fidelity refers to the extent to which the simulation reflects the material aspects (i.e., a physical replica) of the working environment (Lester et al., 2017). It is based on the principle that the more similar the simulated task environment is to the real environment the greater the transfer of learning (Baldwin & Ford, 1998). Psychological fidelity refers to the degree to which the skills and behaviours necessary to complete organisational tasks are accurately represented in the simulated environment (e.g., does the task evoke a similar level of cognitive processing) (Bradley, 2006). Psychological fidelity is expected to elicit similar psychological processes necessary for real-world performance (Kozlowski & DeShon, 2004). The decision on whether to maximise physical fidelity, psychological fidelity, or both during research design is dependent upon the research questions of interest.

Physical fidelity is important when a level of 'dexterity' is needed by the target population to complete the task (Dieckmann et al., 2007). It allows the transfer of procedural skills that might not be possible using psychological fidelity methods alone (Hochmitz & Yuviler-Gavish, 2011), and is especially important when the research question concerns an interplay between humans and hardware (e.g., does a new piece of kit promote faster teamwork?). Understanding the interplay between humans and hardware, referred to as a 'sociotechnical system' (Baxter & Sommerville, 2011), is important for ETs as their context becomes increasingly digitised. ETs where this will be important include control room operators, flight-crews, and emergency medical teams. For example, Stachowski, Kaplan and Waller (2009) used an exact replica of a nuclear control room to study adaptability of teams as they moved through the testing space, communicating and sharing information with colleagues whilst interacting with the electronic displays to rapidly find faults and implement changes to systems (Waller & Kaplan, 2018). Although essential for certain ETs (e.g., operational teams that need to interact with hardware), creating physical fidelity through physical replicas can be difficult as they are often expensive, take up a large amount of physical space, and are often not portable (Kozlowski & De Shon, 2004).

Psychological fidelity is important for researchers interested in studying non-technical skills in ETs (e.g., trust, decision-making, sensemaking), or teams operating at strategic levels. It allows for the examination of the interplay between individual and contextual factors on intra-team processes (Kozlowski & DeShon, 2004). For example, researchers interested in the effects of psychological stressors (e.g., task-related anxiety) on team communication and coordination might build reactionary consequences into the simulation design to increase the gravity of decisions and sense of accountability of decision-makers (Eyre, Crego & Alison, 2008). This might be achieved by gathering team members round a board room style table and providing them with real-time information that follows a realistic narrative to an unfolding situation (e.g., video calls from simulated team members, PDFs with 'data' related to the simulation exercise). An example of where this has been used successfully is Power and Alison (2017b). They ran a simulation study examining how a team of emergency service commanders made decisions during a simulated terrorist incident in which different injects were presented to team members dependent on their answer during the previous inject. This enabled participants to feel immersed by embedding consequences for choices, increasing the gravity of decision-making.

Recent advancements in virtual reality (VR) software offers an accessible and highly immersive way to achieve both physical and psychological fidelity. VR are “computer-generated simulations of three- dimensional objects or environments with seemingly real, direct or physical user interaction” (Dionisio & Gilbert, 2013, p2). They offer an affordable alternative to physical replicas of the organisational environment, whilst still testing important teamwork processes in a context that mirrors the decisions and challenges present in the workplace (Pan & Hamilton, 2018). VR simulations can therefore be used to test both operational (e.g., physical tasks) and strategic teamwork (e.g., decision-making).

One example of a VR system is the Cave Automated Virtual Environment (CAVE). CAVE comprises an enclosed cube, sitting within a large darkened room with projectors on each side (Cruz-Nierra, Sandi, DeFanti, Kenyon, & Hart, 1992). CAVE is attractive as the goggles that are worn do not stop participants from seeing their own hands (as with most head mounted VR devices), whilst they interact with the VR projected on the screens. This means that participants can interact with physical objects (e.g., enact driving by using a real steering wheel) (Pan & Hamilton, 2018), allowing researchers to examine the ability of teams to perform physical tasks. This is especially important when researching ETs that are required to complete arduous physical tasks (e.g., search and rescue teams), and may offer some insight into how contextual demands can influence team members’ ability to use specialist equipment. For example, CAVE has been used to train firefighters using Breathing Apparatus Entry search methods - searching a building for casualties in which sight and breathing is restricted by smoke (Backlund, Engstrom, Hammar, Johannesson & Lebram, 2007). In their study, participants wore personal protective equipment and sensors were fixed to the walls so that physical movements within the “CAVE” corresponded to their movement and orientation within the simulation. This increased the physical effort needed to complete the tasks, giving participants a sense of real-world orientation whilst in a virtual world.

The use of CAVE is not widespread generally (see Jiang, Rahimian, Yon, Plumert, & Kearney, 2016) and this is especially so in relation to ETs. This may be attributed to the fact that it is relatively expensive and difficult to transport in comparison to other VR systems such as head mounted displays (Mallaro, Rahimian, O'Neal, Plumert, & Kearney, 2017). However, evidence from other areas have shown its potential utility for understanding ETs. Gamble et al. (2018) utilised the CAVE system to explore friend/foe discriminatory fire in military personnel, where they found that participants made more errors when under stress, but that 'expertise' was a protective factor. There is also evidence from its use in social psychology that it may be used to explore the role of social influence on individual behaviour. For example, Kinateder et al. (2014) showed that the presence of a virtual agent significantly affected route choice in the evacuation of a tunnel fire. Applied to ETs, the potential for unpacking social influence suggests that CAVE may help develop our understanding in areas such as how intra- and inter- team communication influences performance in multi-team systems (MTS). At present there is limited understanding of how behaviours at the intra-team level affect inter-team performance and vice versa (Asencio & DeChurch, 2017). In the immediate term, and commensurate with the current capability of this kit, we would expect ET research utilising this technology to focus on questions that do not require data to be collected from multiple team members in parallel. In the longer-term, and as this technology advances, we see potential for the CAVE system to study the interaction between multiple individuals, in addition to its current capability of studying the interaction between participants and virtual agents.

When designing a simulation, the involvement of practitioners and/or experts is invaluable. They can ensure that the simulation is relevant to organisational tasks (Klein & Woods, 1993), provide expert input about the task environment and narrative, increase the likelihood that the simulation will elicit similar cognitive and emotional responses found in the real world (Crandall, Klein & Hoffman, 2006), and help to ensure that simulations offer *both* research and training benefits, which can facilitate participant recruitment and engagement (Waller & Kaplan, 2018; Rudolph, Simon & Raemer, 2007). This makes simulations attractive to end-users as they provide a space in which team skills can be trained, facilitating recruitment that might otherwise be challenged by the high workloads and small populations of participants (Beaubien & Baker, 2004).

However, it is important to ensure a balance is met between research and training goals. Simulations can be resource intensive and it is important that researchers are not prevented from collecting the data they need to answer their research questions and practitioners are not promised a simulation that fails to meet their training objectives. To do this, researchers must delineate what the training goals of the organisation are during the early phases of design, and work around them to ensure training objectives are compatible with research goals (Dieckmann et al., 2007). This should facilitate an interdisciplinary partnership and enable collaboration through the entire research project. The involvement of practitioners at the early stages of research can also have benefits later on in terms of research dissemination and impact. Practitioners are keen to receive feedback on their training, as such, a research team might want to organise a feedback workshop or write a practitioner-friendly report on findings. This can facilitate opportunities for further follow-up studies and ensure a collaborative relationship with practitioners moving forward.

Data collection

A key benefit to simulation research is that it facilitates the collection of rich behavioural data, allowing researchers to study the verbal and non-verbal dynamics of teamwork. Psychology has seen a decline in the use of behavioural measures in recent years, typically showing a tendency to use self-report surveys (Cialdini, 2009; Dolinski, 2018).

However, there has been a general call to move beyond self-report measures to gain a better understanding of how social coordination emerges in complex environments (Willemsen-Dunlap et al., 2018) and to develop more objective measures of behaviour (Rosen & Dietz 2017). This is due, in part, to the limitations of solely using self-report measures which; (i) fail to account for the richness of team-based interactions (Shuffler & Carter, 2018); (ii) lead to a proliferation of scales each attempting to measure the same thing (see Salas et al., 2015 on team cohesion); (iii) show weak correspondence with non self-report outcome measures (*see* Valentine, Nembhard & Edmonson, 2015 for a review in a health care setting), and (iv) are subject to a number of biases (e.g., Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). We suggest that simulations offer a methodological advantage to self-report by recording behaviour in situ.

Wearable technology. The tools used to collect data during simulations need to be unobtrusive so as not to break immersion, but robust enough to allow for reliable examination of the research question. The advancement of behavioural measures creates promise for the use of wearable sensors in research using simulations. Wearable sensors are mobile devices that record data on how the wearer interacts with their surroundings (including other people). They do this using microphones, accelerometers, infrared sensors and/or Bluetooth components (Chaffin et al., 2017). Wearable sensors have advantages over traditional methods; namely that they allow for the effortless recording of data from participants that are not reliant on self-reports, and that data are real-time and continuously collected thus removing the necessity for researchers to piece together static data taken at set times, sometimes from multiple devices. This makes wearable sensors especially suited to simulations, as the continuous collection of rich data in the real world may lead to consent and confidentiality issues (e.g., recording patient-clinician interactions).

The fact that behavioural data are collected continuously means that wearable devices have the potential to identify important *within*-person insights and their impact on team performance. This has not always been achieved with traditional methods, which tend to focus at the *between*-person level (Matusik et al., 2019). This finer grained understanding of how teams operate has the potential for simulation methods to develop complex, non-linear, relationships between relational variables. For example, data from wearable sensors may allow for the development of a finer-grained understanding of leadership in ETs, such as how a leader's behaviour fluctuates across an emergency and how these fluctuations impact behaviours. Similarly, it may examine how leadership changes interact with team factors (e.g., the presence of other teams – as within MTS) or external forces (e.g., contextual demands – during crises response).

At a theoretical level, wearable sensors are most valuable when the research question concerns relational issues at the team level (e.g., cohesion, trust, leadership), as they show how the person navigates their environment, including social interactions. In using data from single or multiple streams (e.g, audio, Bluetooth), studies have used wearable technology to examine affect and team cohesion in simulated space exploration missions (Zhang et al., 2018), cooperation (Taylor, 2013), communication in productive and creative teams (Pentland, 2012), social and task-related exchanges (Matusik et al., 2019), social networks (Wu, Waber, Aral, Brynjolfsson & Pentland, 2008), boundary spanning individuals (i.e., those that coordinate activity between established groups) (Chaffin et al. (2017), and emergent leaders (Chaffin et al., 2017). There is potential for research in ETs to build on this to use sensors in the study of MTS, to explore how boundary spanning individuals support teamworking across multiple agencies responding in crises. Previous research has tended to rely on self-report and coding of verbal behaviours (*see* Bienefeld & Grote, 2014), whereas wearables can measure other aspects such as variations in proximity over time (i.e., using Bluetooth), in addition to providing a continuous measure of communication.

Research using audio data more generally expands the potential for wearables in simulation research. For example, Stanton and Roberts (2018) used audio data to understand team level macrocognition (i.e., cognitive functions that are performed in naturalistic settings, see Klein, Ross, Moon, Klein, Hoffman & Hollnagel, 2003), Bowers, Jentsch, Salas and Braun (1998) have used it to understand shared mental models, and Fischer, Donnel and Orasanu (2007) have used it to examine which *types* of information (task or relational focused) best support performance in ETs. From the perspective of understanding ETs, this is especially promising as the nature of these environments means that team members have to share, analyse and discuss complex information (e.g., Haddow & Bullock, 2003). An important question for ETs, due to the time sensitive nature of their work, is how to do this efficiently. Evidence from a range of non-ETs suggests that short and equal verbal contributions, face-to-face communication, distributed connections within the team and information seeking from other teams characterise success (Pentland, 2012). Wearable sensors would allow for a reliable test of these hypothesised effects in ETs, whilst maintaining the realism of the ET environment through the use of the simulation.

Wearable technology can be used to record physiological data from team members. Psychological pressures (e.g., stress) is an important factor to consider in simulation research of ETs as the inherently stressful environments they operate in can disrupt performance (Driskell et al., 2018). Stress has been shown to reduce communication in aviation teams (Sexton & Helmreich, 2000), impair cognitive functioning in military teams (Wallenius, Larrson & Johannsson, 2004) and reduce information sharing in less experienced surgical teams (Wetzel et al., 2006). Although previous research has explored the role of stress in ETs, studies have often failed to check whether the experimental manipulation has actually affected stress levels or, alternatively, used a self-report survey to do so. For example, increasing stress by imposing time pressure has been associated with an increase in risk-taking behaviour (Young, Goodie, Hall & Wu, 2012), and a shift towards more satisficing decision styles (Alison, Doran, Long, Power & Humphrey, 2013). However, neither of these studies took physiological measures of stress from their participants and so the effects of stress, via time pressure, were assumed.

Wearable technology allows us to address the limitation of these other studies. It is possible to measure stress during a simulation by using wearables that record ‘stress-related’ measures, such as heart-rate, galvanic skin responses and change in pitch (Mozos et al., 2017). For example, stress during a simulated driving task, as measured using skin conductivity (i.e., sweating) and heart rate, has been found to predict stress levels with the highest level of accuracy when compared against physical indicators of stress (e.g., breaking and sharp turning) and self-report measures (Healey & Picard, 2005). Heart rate has also been identified as the best indicator of stress in a study comparing physiological indicators of stress during a simulated virtual environment that invoked fear by placing participants over a chasm at great height (Meehan, Insko, Whitton & Brooks Jr, 2002). When applied to ETs, physiological indicators of stress open the possibility of building models that map team responses across a stress episode: from its origin through peak to end. What sets these models apart from conventional teams (where such devices are equally insightful) is the potential for ET models to overlay the stress episodes experienced by inter-related teams (e.g., MTS) to examine interplay or contagion.

In keeping with the need to maintain fidelity during simulations, researchers may also consider using physiological measures of stress to provide an objective indication of how immersive a simulation has been. Baker et al. (2017) used a heart rate monitor to assess if the stress experienced in medical procedures could be replicated within a simulated environment and found that the simulated procedure did not accurately re-create the same level of stress as experienced within hospitals. This emphasises the need to incorporate a physiological measure of stress to ensure that elements of the simulation that are intended to be difficult induce a level of urgency within the participants. There is currently a lack of research that has sought to establish what stress levels are needed to ensure that simulations are useful for training and research purposes (Cumin, Boyd, Webster & Weller 2013). More research is needed to establish standardized levels of immersion which can leave organisations confident that simulations are achieving their intended purposes (Cumin, Weller, Hender & Merry, 2010).

Interactions within the simulation system. Although wearable sensors have the potential to provide rich data on relational issues at the team level, they may not be able to provide a holistic overview of teamwork, such as when communication occurs via other mediums (e.g., email) or when inter-dependent tasks are carried out in different locations. For example, some ETs (e.g., MTS) will operate across several sites and researchers may wish to explore how cultural factors (e.g., organisational policies) and team structures facilitate/hinder inter-team processes. One benefit of simulations is that teams are operating in designated room(s), and so forms of data collection can be built into the simulation system to provide a comprehensive account of verbal and non-verbal communication between team members. Data gathered from participant interactions within the simulation system might include video recording, for example, CCTV of the team operating in the simulation room; or recording data within the simulation computer system itself, for example, by generating a log of clicks or button pushes when participants interact with the simulation; collecting time-stamped 'decision logs'; and eye tracking on the computer screen. Monitoring the interaction within the simulation system may prove particularly important for researchers interested in designing a simulation with high physical fidelity to explore sociotechnical systems (e.g., how team members interact with the computer system). Future research could consider how simulations with high physical fidelity might advance theory on sociotechnical systems and their use by ETs. For example, in considering the role of the team in increasingly automated systems or in what way do contextual demands (e.g., dynamic task requirements) impact team members' ability to effectively utilise technology in crises.

The type of data recorded in the simulation system will be dependent on the system being used and the research questions of interest. For example, research questions that are interested in how team-level factors (e.g., composition) influence decision speed might use a time-stamped 'decision log'. Power and Alison (2017b) used this method to identify how long it took teams to make decisions and how this interacted with the team's goal. Teams were requested to 'log' their decisions on a computer when they wanted to make a decision and these data were automatically recorded and timestamped in the simulation system. Alternatively, researchers may use the simulation system to monitor how team members communicate electronically with one another. Alison et al. (2015) were interested in communication patterns between sub-teams in different 'syndicate' rooms in a simulation. To do this, they built a 'chatbox' function into the simulation system so that sub-teams could communicate between rooms, with all electronic communications data recorded and time stamped. The simulation system therefore offers an alternative mode of data collection that can be used in isolation or in conjunction with wearable devices dependent on the research question.

Data Analysis

Simulation research with ETs has the potential to yield vast amounts of data from multiple sources, measuring multiple variables. It is important that data analysis maximises understanding of this rich data. There exists a number of methods of analysis that can be used. Here, we focus on two types that are especially relevant: (i) network analyses, which examine interpersonal dynamics within a team at a single time point; and (ii) temporal analyses, which track interpersonal dynamics over time. We focus on these methods as they provide rich representations of *team interactions*, as oppose to assessing the *individual* performance of team-based skills (e.g., Yule et al., 2008). We then turn our attention to the possibility of using Bayesian statistics, which allow analyses to be carried out with smaller samples, and thus may open up the possibility of testing more complex theoretical models in ET research.

Network analyses. Network analyses allow a researcher to analyse team behaviour during simulations by quantifying information and providing a visual representation of how team members interact. This type of analysis is especially useful when comparing how contextual factors (e.g., task type) influence team behaviours (e.g., inter-team communication) (Stanton & Roberts, 2018). Using recorded communication data (e.g., by using wearable devices or CCTV recordings), Social Network analysis (SNA) shows how team members communicate with each other and the centrality of any one member (Knoke & Yang, 2008). SNA are also useful as they provide a visual representation of the social dynamics of a team by plotting each person as a node and showing the strength of the connections between them. At a theoretical level, this is especially important for ETs that involve multiple agencies operating within a hierarchical structure as it can identify instances in which communication patterns do not follow pre-defined organisational processes and structures (Dekker, 2000), or plausible reasons for communication breakdowns. For example, SNA has been used to identify key tasks that challenged communication in submariners (Stanton & Roberts, 2018), how team communications varied dependent on team composition in surgical operating staff (Anderson & Talsma, 2011), and how a lack of connectedness between a Search and Rescue Team contributed to faulty communications and the ability to develop shared situation awareness (Fodor & Flestea, 2016).

An alternative type of network analysis that goes beyond communications data is the Event Analysis of Systematic Teamwork (EAST) technique. This method models the macrocognition (i.e., situation awareness) of a team by generating task and information networks in addition to social networks (Walker, Gibson, Stanton, Baber, Salmon & Green, 2006). In order to perform EAST, raw data from audio and video recordings are transcribed and then used to create matrices of each of the three networks (i.e., social, task, information). This results in a “network of networks”, that allows researchers to identify how constructs in different networks might interrelate. For example, communications might influence the way a task is performed, which might influence how information is transferred.

EAST has been used to examine teamwork in simulation research across several extreme contexts; submariner command and control (Stanton & Roberts, 2018); emergency response (Houghton, Baber, McMaster, Salmon, Stewart & Walker, 2006) and air traffic control (Walker et al., 2006). As EAST involves generating a task network, it is useful for researchers who are interested in understanding how team members coordinate to complete tasks as well as how they communicate with one another in extreme environments. Hierarchical Task Analysis is a methodology within EAST that is used to identify key tasks (Annett & Stanton, 2000), as well as the individuals who complete tasks, the structure, and the order in which the tasks take place (Walker et al., 2006). This provides a detailed representation of how team goals interact and are resolved (Walker, Stanton, Baber, Wells, Gibson, Salmon, & Jenkins, 2010). For example, a simulation researcher interested in team coordination may want to model how a team approaches different tasks dependent on difficulty. As coordination is defined as the behavioural mechanism enabling teams to sequence, synchronize and integrate their efforts in order to achieve goal-relevant tasks

(Marks et al., 2001), modelling how teams move through tasks should contribute to a more complex understanding of how ETs coordinate. This is extremely relevant for researchers interested in ETs due to the importance of coordination in manging complex team structures and preventing error across a range of contexts such as aviation (Grote, Kolbe, Zala-Mezo, Bienefeld-Seall & Kunzle, 2010) and medical emergency teams (Schmutz & Manser, 2013).

Temporal analysis. Temporal analysis seeks to identify how team behaviour might change over time in response to changes in individual, team and contextual demands. This type of analysis is especially useful for ET researchers interested in exploring how team processes emerge and are sustained during simulated tasks. It recognises the important role of context in shaping team-based interactions (Ilgen, 1999), emphasising that teamwork does not exist in a vacuum and team processes will change over time (Kozlowski & Ilgen, 2006). Non-simulation based team research has sought to study how teamwork changes over time by collecting longitudinal data (e.g., questionnaires) at set intervals over a given period (see, Mathieu et al., 2015). However, this staged approach might not be feasible for some ETs as team members rotate and might not work together at set regular intervals (e.g., emergency response teams). Moreover, these approaches tend to rely on self-report data, as oppose to monitoring actual behaviour in real-time, which has limitations as detailed above (Shuffler & Carter, 2018).

An alternative approach is to study how team behaviour evolves during a simulation. Although simulations will not produce ‘longitudinal’ temporal data in the traditional sense (e.g., over a course of weeks/months/years), simulations offer a closer replica of how ETs operate in the real world, wherein they must adapt and evolve their teamwork during a given task (e.g., emergency incident). As such, simulations allow us to study the temporal dynamics of teamwork during a simulated ‘event’, which can incorporate multiple goal directed tasks and episodes (Marks, Mathieu & Zaccaro, 2001). By analysing simulation data longitudinally (i.e. over the course of the simulation), researchers can explore how teams adapt and change as they cycle through different episodes within the simulated event (Marks et al., 2001). The advent of wearables and advancements here allows for this to be done in a reliable and highly detailed way, enabling researchers to begin examining complex, non-static theories or models of behaviour. This could be especially important to advance understanding of MTS. For example, wearable devices may be used to measure communication and relational emerging variables such as cohesion across multiple component teams. When coupled with repeated SNA this would allow researchers to map how intra- and inter- team behaviours and relationships change over time. This could answer questions such as how intra-team behaviours relate to inter-team performance or how intra-team cohesion affects how inter-team members relate to one another.

Beyond comparing networks analyses during different phases of a simulation, a more complex way of analysing temporal data is by using lag sequential analyses, which seek to identify non-random patterns of behaviour during a task (Becker-Beck, 2001). It is useful for research questions that seek to identify how specific team behaviours (e.g., shifts in communication patterns across team members) develop and change over time (Leenders, Contractor, & DeChurch, 2017) and how specific patterns of behaviour can lead to better team performance (Kauffeld & Meyers, 2009). An example of how lag sequential analyses have been used to study ETs during simulations is Cohen-Hatton, Butler and Honey (2015). Their research sought to identify whether commanders in the Fire Service prescribed to the standard decision model used by the Fire Service, or whether they deviated. Participants were asked to "think aloud" (i.e., verbalise their thoughts) during a simulation, and transcripts were coded to identify if participants progressed through the prescribed model of "situation assessment" to "plan formulation" to "plan execution". They found, using lag-sequential analyses, that participants did not follow this pattern. However, a simple goal-oriented training intervention made participants more likely to adopt the prescribed processing pattern, without delaying decision speed. Lag sequential analyses are thus useful for helping to understand patterns in team processing and behaviour during a simulated event, and also provides possibilities for testing interventions to increase adherence to decision models and/or improve performance. For example, using this technique, research might develop our understanding of how patterns of behaviours change dependent on information flow, level of stress in team members (as measured using physiological markers), changes in goal hierarchies; and the interaction between these variables. In doing so, we would have an enhanced understanding of the temporal and contextual influences on teamwork in ETs.

Bayesian statistics. Another approach to analysing data from ET simulations is by using Bayesian statistics. Unlike network and temporal analyses, Bayesian statistics are not a type of data analysis, but are an alternative statistical approach to classic significance testing. Traditional research on teams often draws on classic significance testing (e.g., null hypothesis testing, p-values, confidence intervals) to test specific variables and theories. However, this approach is problematic when working with ETs as, at a practical level, it often calls for moderate to large sample sizes with normal distributions (see Wagenmakers et al., 2018 for other problems with classic theory). Research with ETs tends to involve small sample sizes as the participant pool is much smaller than the general population and participants often have limited time to take part in research (Bell et al., 2018). Whilst efforts to address this have drawn on using trainees from ETs, such as trainee paramedics (e.g., Amacher et al., 2017), these samples have been shown to operate differently to 'experts' (Boulton & Cole, 2016). In other types of ETs (e.g., emergency response, command and control), trainees may also not be as readily available as they are in clinical settings.

In response to problems with classic testing, researchers are calling for alternative methods of analysis (e.g., Vandekerckhove, Rouder, & Kruschke, 2018). One that has seen an increase in popularity—facilitated by advancements in computer algorithms and quicker hardware processing—are Bayesian Statistics (for example, see Special Issues in *Journal of Mathematical Psychology* 2016, vol. 72; *Psychonomic Bulletin & Review* 2018; vol. 25). As a set of tools, Bayesian statistics are attractive to ET simulation research as they open the potential, *inter alia*, for theoretical models to be tested even when samples are smaller than with conventional team research.

As a very broad (and somewhat simplified) overview, Bayesian statistics have the ultimate goal of showing the probability that the data observed is likely to occur under two competing theoretical (i.e., statistical) models (Kruschke & Liddell, 2018). Using Bayes factors, a researcher infers the level of support for their theory, relative to the alternative theory, based on how much the observed data differs from that predicted. This is done by comparing the statistical model against a 'posterior' probability distribution, which is made up from prior information known before data were collected and what is known from the actual – observed – data. Prior knowledge can come from theoretical frameworks, findings from previous research, subject experts and pilot work (Zyphur & Oswald, 2015). Research may also use non-informative priors where knowledge is limited and parameters are set to cover a broad range of possible outcomes, but this is less advisable when samples are small (see, McNeish, 2016). Bayesian statistics regard parameters (e.g., probabilities) as variables, and as such, parameters are adjusted as data accumulates and output is compared against starting values. The researcher can thus see how evidence for their theoretical (statistical) model changes with new data; something that is not possible with classic theory where parameters are regarded as constant (see, Gelman et al., 2014 for a statistical overview of Bayes analysis; Lynch, 2010 for a general introduction; Jeffrey, 1961 for original writings).

Classical significance tests require researchers to specify in advance what the smallest effect size of interest is given their theory in order to recruit a sufficient number of participants capable of detecting such an effect. Yet, it has been shown using Bayesian analyses that a high powered non-significant result might not necessarily constitute evidence for the null, and that low-powered non-significant results are not necessarily insensitive (Dienes, & McLatchie, 2018). Evidence suggests that sample sizes estimated using parameters generated through Bayesian analysis rather than power, may be more flexible and yield smaller sample size requirements (Sambucini, 2017). Relatedly, Bayesian analysis has the benefit of allowing for ‘optimal stopping’. In essence, this allows for a researcher to track results as data are collected and stop data collection when a certain level of evidence showing one theory as more favourable has been obtained (Kelley, 2013). In addition to allowing for potentially smaller samples to be tested to obtain an effect, this also avoids the ethical issue of testing ET members beyond what is needed.

Bayesian analyses have been applied to a number of methods from t-tests through to structural equation modelling (Brown, Barrett & Power, 2019; McNeish, 2016). For ETs, it could be applied to existing methods (e.g., using a t-test to compare two sets of SNA across phases of a simulation) to identify significant effects that may have been masked by small sample sizes. Depending on the complexity of the theoretical model, we may start to move towards unpacking the different pathways through which factors have an effect on team performance, and the conditions that moderate these effects. This could be especially important in understanding the complex inter-play between component team and system level variables in MTS, which linear approaches may not be able to account for (Cronin, 2015). As interest in larger multi-agency teams expands, we may see the use of Bayesian methods grow as researchers seek to test theoretical frameworks that span multiple levels (i.e., variables at the component team and system level) that traditional statistical approaches would not have the power to do when working with small sample sizes (Wang & Hanges, 2011).

Conclusion

Teamwork is a necessity in almost any twenty-first century organisation, with teams increasingly viewed as the solution to solving complex problems (Salas, Shuffler, Thayer, Bedwell & Lazarra, 2015). This is especially so in organisations operating in extreme environments, where team members must coordinate their behaviour effectively in order to avoid the severe, often life or death, consequences of poor performance. In this paper we have identified the benefits to conducting simulation research with ETs, showing how they differ from existing methods. Second, we have presented a framework for conducting immersive simulations, focusing on three broad aspects of; (i) study design, (ii) data collection and (iii) data analysis. By doing this we have reviewed existing simulation research, as well as suggesting how emerging technologies (e.g., wearable devices, CAVE) and statistical methods (e.g., Bayesian) might be used in simulation research to advance understanding. It is hoped that this paper will inspire researchers to make use of novel immersive simulation-based methods to engender the much-needed empirical research on ETs.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This work was funded by the Centre for Research and Evidence on Security Threats [ESRC Award: ES/N009614/1].

References

Alison, L., Doran, B., Long, M. L., Power, N., & Humphrey, A. (2013). The effects of subjective time pressure and individual differences on hypotheses generation and action prioritization in police investigations. *Journal of Experimental Psychology: Applied*, *19*(1), 83.

Alison, L., Power, N., van den Heuvel, C., Humann, M., Palasinksi, M., & Crego, J. (2015). Decision inertia: Deciding between least worst outcomes in emergency responses to disasters. *Journal of Occupational and Organizational Psychology*, *88*(2), 295-321.

Alison, L., van den Heuvel, C., Waring, S., Power, N., Long, A., O'Hara, T., & Crego, J. (2013). Immersive simulated learning environments for researching critical incidents: A knowledge synthesis of the literature and experiences of studying high-risk strategic decision making. *Journal of Cognitive Engineering and Decision Making*, *7*(3), 255-272.

Amacher, S. A., Schumacher, C., Legeret, C., Tschan, F., Semmer, N. K., Marsch, S., & Hunziker, S. (2017). Influence of gender on the performance of cardiopulmonary rescue teams: a randomized, prospective simulator study. *Critical Care Medicine, 45*(7), 1184-1191.

Anderson, C., & Talsma, A. (2011). Characterizing the structure of operating room staffing using social network analysis. *Nursing Research*, *60*(6), 378-385.

Annett, J., & Stanton, N. A. (Eds.). (2000). *Task analysis*. London/New York. CRC Press.

Asencio, R., & DeChurch, L. A. (2017). Assessing collaboration within and between teams: A multiteam systems perspective. In A. von Davier, M. Zhu, & P. Kyllonen (Eds.), *Innovative assessment of collaboration. Methodology of educational measurement and assessment*. Cham: Springer.

Backlund, P., Engstrom, H., Hammar, C., Johannesson, M., & Lebram, M. (2007, July). Sidh-a game based firefighter training simulation. In *Information Visualization, 2007. IV'07. 11th International Conference* (pp. 899-907). IEEE.

Baker, B. G., Bhalla, A., Doleman, B., Yarnold, E., Simons, S., Lund, J. N., & Williams, J. P. (2017). Simulation fails to replicate stress in trainees performing a technical procedure in the clinical environment. *Medical Teacher*, *39*(1), 53-57.

Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology*, *41*(1), 63-105.

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers, 23*(1), 4-17.

Beaubien, J. M., & Baker, D. P. (2004). The use of simulation for training teamwork skills in health care: how low can you go?. *BMJ Quality & Safety*, *13*(1), i51-i56.

Becker-Beck, U. (2001). Methods for diagnosing interaction strategies: An application to group interaction in conflict situations. *Small Group Research, 32*(3), 259-282.

Bell, S. T., Fisher, D. M., Brown, S. G., & Mann, K. E. (2018). An approach for conducting actionable research with extreme teams. *Journal of Management*, *42*(5)*,* 1-26.

Bienefeld, N., & Grote, G. (2014). Shared leadership in multiteam systems: How cockpit and cabin crews lead each other to safety. *Human Factors*, *56*(2), 270-286.

Boulton, L., & Cole, J. (2016). Adaptive flexibility: Examining the role of expertise in the decision making of authorized firearms officers during armed confrontation. *Journal of Cognitive Engineering and Decision Making*, *10*(3), 291-308.

Bowers, C. A., Jentsch, F., Salas, E., & Braun, C. C. (1998). Analyzing communication sequences for team training needs assessment. *Human Factors, 40*(4), 672-679.

Bradley, P. (2006). The history of simulation in medical education and possible future directions. *Medical Education*, *40*(3), 254-262.

Brown, O., Barrett, E., Power, N. (2019) Monitoring cohesion over time in expedition teams; the role of daily events and team composition. In *Proceedings of the 14th International Naturalistic Decision-Making Conference, San Francisco, U.S.A*.

Chaffin, D., Heidl, R., Hollenbeck, J. R., Howe, M., Yu, A., Voorhees, C., & Calantone, R. (2017). The promise and perils of wearable sensors in organizational research. *Organizational Research Methods*, *20*(1), 3-31.

Cheng, A., Eppich, W., Grant, V., Sherbino, J., Zendejas, B., & Cook, D. A. (2014). Debriefing for technology-enhanced simulation: a systematic review and meta-analysis. *Medical Education, 48*(7), 657-666.

Cheng, A., Kessler, D., Mackinnon, R., Chang, T. P., Nadkarni, V. M., Hunt, E. A., ... & Hui, J. (2016). Reporting guidelines for health care simulation research: extensions to the CONSORT and STROBE statements. *Advances in Simulation*, *1*(1), 25.

Cialdini, R. B. (2009). We have to break up. *Perspectives on Psychological Science*, *4*(1), 5-6.

Cohen-Hatton, S. R., & Honey, R. C. (2015). Goal-oriented training affects decision-making processes in virtual and simulated fire and rescue environments. *Journal of Experimental Psychology: Applied*, *21*(4), 395.

Crandall, B., Klein, G., Klein, G. A., & Hoffman, R. R. (2006). *Working minds: A practitioner's guide to cognitive task analysis*. Mit Press.Cross, R., Rebele, R., & Grant, A. (2016). Collaborative overload. *Harvard Business Review*, *94*(1).

Crichton, M. T., Flin, R., & Rattray, W. A. (2000). Training decision makers–tactical decision games. *Journal of Contingencies and Crisis Management*, *8*(4), 208-217.

Cronin, M. A. (2015). Advancing the science of dynamics in groups and teams. *Organizational Psychology Review, 5* (4), 267–269.

Cruz-Neira, C., Sandin, D. J., DeFanti, T. A., Kenyon, R. V., & Hart, J. C. (1992). The CAVE: audio visual experience automatic virtual environment. *Communications of the ACM*, *35*(6), 64-73.

Cumin, D., Boyd, M. J., Webster, C. S., & Weller, J. M. (2013). A systematic review of simulation for multidisciplinary team training in operating rooms. *Simulation in Healthcare*, *8*(3), 171-179.

Cumin, D., Weller, J. M., Henderson, K., & Merry, A. F. (2010). Standards for simulation in anaesthesia: creating confidence in the tools. *British Journal of Anaesthesia*, *105*(1), 45-51.

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science*, *323*(5910), 66-69.

Dekker, A. H. (2000, October). Social network analysis in military headquarters using CAVALIER. In *Proceedings of Fifth International Command and Control Research and Technology Symposium* (pp. 24-26).

Dieckmann, P., Gaba, D., & Rall, M. (2007). Deepening the theoretical foundations of patient simulation as social practice. *Simulation in Healthcare*, *2*(3), 183-193.

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, *25*(1), 207-218.

Dionisio, J. D. N., & Gilbert, R. (2013). 3D Virtual worlds and the metaverse: Current status and future possibilities. *ACM Computing Surveys (CSUR)*, *45*(3), 34.

Doliński, D. (2018). Is psychology still a science of behaviour? *Social Psychological Bulletin*, *13*(2), 1-14.

Driskell, T., Salas, E., & Driskell, J. E. (2018). Teams in extreme environments: Alterations in team development and teamwork. *Human Resource Management Review*. Human Resource Management Review. 28(4), 434-449.

Dwyer, T., & Raftery, A. E. (1991). Industrial accidents are produced by social relations of work: A sociological theory of industrial accidents. *Applied ergonomics, 22*(3), 167-178.

Eyre, M., Crego, J., & Alison, L. (2008). Electronic debriefs and simulations as descriptive methods for defining the critical incident landscape. In L. Alison & J. Crego (Eds.), *Policing critical incidents* (pp. 24–53). Cullumpton, UK: Willan.

Fischer, U., McDonnell, L., & Orasanu, J. (2007). Linguistic correlates of team performance: Toward a tool for monitoring team functioning during space missions. Aviation, Space, and *Environmental Medicine, 78*(5), B86-B95

Fodor, O. C., & Flestea, A. M. (2016). When fluid structures fail: A social network approach to multi-team systems' effectiveness. *Team Performance Management*, *22*(3/4), 156-180.

Gamble, K. R., Vettel, J. M., Patton, D. J., Eddy, M. D., Davis, F. C., Garcia, J. O., ... & Brooks, J. R. (2018). Different profiles of decision making and physiology under varying levels of stress in trained military personnel. *International Journal of Psychophysiology*, *131*, 73-80.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997-1016.

Gillespie, B. M., Gwinner, K., Chaboyer, W., & Fairweather, N. (2013). Team communications in surgery–creating a culture of safety. *Journal of Interprofessional Care*, *27*(5), 387-393.

Grote, G., Kolbe, M., Zala-Mezö, E., Bienefeld-Seall, N., & Künzle, B. (2010). Adaptive coordination and heedfulness make better cockpit crews. *Ergonomics*, *53*(2), 211-228.

Haddow, G. D., & Bullock J, A. (2003): Introduction to Emergency Management. MA, USA: Butterworth Heinemann.

Healey, J., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, *6*(2), 156-166.

Hochmitz, I., & Yuviler-Gavish, N. (2011). Physical fidelity versus cognitive fidelity training in procedural skills acquisition. *Human Factors, 53*(5), 489-501.

Houghton, R. J., Baber, C., McMaster, R., Stanton, N. A., Salmon, P., Stewart, R., & Walker, G. (2006). Command and control in emergency services operations: a social network analysis. *Ergonomics*, *49*(12-13), 1204-1225.

Hughes, A. M., Gregory, M. E., Joseph, D. L., Sonesh, S. C., Marlow, S. L., Lacerenza, C. N., ... & Salas, E. (2016). Saving lives: A meta-analysis of team training in healthcare. *Journal of Applied Psychology*, *101*(9), 1266.

Ilgen, D. R. (1999). Teams embedded in organizations: Some implications. *American Psychologist*, *54*(2), 129.

Jeffreys, H. (1961). Theory of probability. Oxford: Oxford University Press, Clarendon Press

Jiang, Y., O'Neal, E., Rahimian, P., Yon, J. P., Plumert, J. M., & Kearney, J. K. (2016). Action coordination with agents: crossing roads with a computer-generated character in a virtual environment. In *Proceedings of the ACM Symposium on Applied Perception* (pp. 57-64). ACM.

Kadiri, Z. O., Nden, T., Avre, G. K., Oladipo, T. O., Edom, A., Samuel, P. O., & Ananso, G. N. (2014). Causes and Effects of Accidents on Construction Sites (A Case Study of Some Selected Construction Firms in Abuja, FCT Nigeria). *IOSR Journal of Mechanical and Civil Engineering*, *11*(5), 66-72.

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*(3), 267-294.

Kelley, K. (2013). Effect size and sample size planning. In Little, T.D. (Ed.) Oxford Handbook of Quantitative Methods (pp. 206–222). New York: Oxford University Press.

Kinateder, M., Ronchi, E., Gromer, D., Müller, M., Jost, M., Nehfischer, M., ... & Pauli, P. (2014). Social influence on route choice in a virtual reality tunnel fire. *Transportation Research Part F: Traffic Psychology and Behaviour*, *26*, 116-125.

Klein, G. A., & Woods, D. D. (1993). Conclusions: Decision making in action. In G.A. Klein, J. Orasanu, R. Calderwood & C.E. Zsambok (Eds.) *Decision making in action: Models and methods*, (pp. 404-411). Ablex

Klein, G., Ross, K. G., Moon, B. M., Klein, D. E., Hoffman, R. R., & Hollnagel, E. (2003). Macrocognition. *IEEE Intelligent Systems*, *18*(3), 81-85.

Klein, R. L., Bigley, G. A., & Roberts, K. H. (1995). Organizational culture in high reliability organizations: An extension. *Human Relations*, *48*(7), 771-793.

Klein, K. J., Ziegert, J. C., Knight, A. P., & Xiao, Y. (2006). Dynamic delegation: Shared, hierarchical, and deindividualized leadership in extreme action teams. *Administrative Science Quarterly, 51*(4), 590-621.

Knoke, D., & Yang, S. (2008). *Social Network Analysis*. Sage.

Kozlowski, S. W. (2015). Advancing research on team process dynamics: Theoretical, methodological, and measurement considerations. *Organizational Psychology Review*, *5*(4), 270-299.

Kozlowski, S. W., & DeShon, R. P. (2004). A psychological fidelity approach to simulation-based training: Theory, research and principles. *Scaled worlds: Development, validation, and applications*, 75-99.

Kozlowski, S. W., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest*, *7*(3), 77-124.

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178-206.

Leenders, R. T. A., Contractor, N. S., & DeChurch, L. A. (2016). Once upon a time: Understanding team processes as relational event networks. *Organizational Psychology Review, 6*(1), 92-115.

Lester, E., Georgiou, A., Hein, M., Littlepage, G., Moffett III, R., & Craig, P. (2017). Improving Aviation Students' Teamwork, Problem Solving, Coordination, and Communications Skills During a High-Fidelity Simulation. In *19th International Symposium on Aviation Psychology* (p. 119).

Lynch, S.M. (2010). Introduction to applied Bayesian statistics and estimation for social scientists. New York: Springer.

Mallaro, S., Rahimian, P., O'Neal, E. E., Plumert, J. M., & Kearney, J. K. (2017). A comparison of head-mounted displays vs. large-screen displays for an interactive pedestrian simulator. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology (pp. 1-4).*

Manser, T., Dieckmann, P., Wehner, T., & Rall, M. (2007). Comparison of anaesthetists' activity patterns in the operating room and during simulation. *Ergonomics*, *50*(2), 246-260.

Manzey, D., & Lorenz, B. (1998). Mental performance during short-term and long-term spaceflight. *Brain Research Reviews*, *28*(1-2), 215-221.

Maran, N. J., & Glavin, R. J. (2003). Low-to high-fidelity simulation–a continuum of medical education? *Medical Education*, *37*, 22-28.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, *26*(3), 356-376.

Mazzocco, K., Petitti, D. B., Fong, K. T., Bonacum, D., Brookey, J., Graham, S., ... & Thomas, E. J. (2009). Surgical team behaviors and patient outcomes. *The American Journal of Surgery*, *197*(5), 678-685.

McCabe, B., Loughlin, C., Munteanu, R., Tucker, S., & Lam, A. (2008). Individual safety and health outcomes in the construction industry. *Canadian Journal of Civil Engineering*, *35*(12), 1455-1467.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modelling: A Multidisciplinary Journal*, *23*(5), 750-773.

Matusik, J. G., Heidl, R., Hollenbeck, J. R., Yu, A., Lee, H. W., & Howe, M. (2018). Wearable bluetooth sensors for capturing relational variables and temporal variability in relationships: A construct validation study. *Journal of Applied Psychology*, *104*(3), 357-387.

Maynard, M. T., Kennedy, D. M., & Resick, C. J. (2018). Teamwork in extreme environments: lessons, challenges, and opportunities. *Journal of Organizational Behaviour*, *39*(6), 695-700.

Meehan, M., Insko, B., Whitton, M., & Brooks Jr, F. P. (2002). Physiological measures of presence in stressful virtual environments. *ACM Transactions on Graphics (tog)*, *21*(3), 645-652.

Mozos, O. M., Sandulescu, V., Andrews, S., Ellis, D., Bellotto, N., Dobrescu, R., & Ferrandez, J. M. (2017). Stress detection using wearable physiological and sociometric sensors. *International Journal of Neural Systems*, *27*(2), 1-17.

Pan, X., & Hamilton, A. F. D. C. (2018). Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology*, *109*(3), 395-417.

Pentland, A. (2012). The new science of building great teams. *Harvard Business Review*, *90*(4), 60-69.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879.

Power, N., & Alison, L. (2017a). Redundant deliberation about negative consequences: Decision inertia in emergency responders. *Psychology, Public Policy, and Law*, *23*(2), 243.

Power, N., & Alison, L. (2017b). Offence or defence? Approach and avoid goals in the multi-agency emergency response to a simulated terrorism attack. *Journal of Occupational and Organizational Psychology*, *90*(1), 51-76.

Risser, D. T., Rice, M. M., Salisbury, M. L., Simon, R., Jay, G. D., Berns, S. D., & MedTeams Research Consortium. (1999). The potential for improved teamwork to reduce medical errors in the emergency department. *Annals of Emergency Medicine*, *34*(3), 373-383.

Roberts, K. H. (1990). Managing high reliability organizations. *California management review*, *32*(4), 101-113.

Roma, P. G., & Bedwell, W. L. (2017). Key factors and threats to team dynamics in long-duration extreme environments. In *Team Dynamics Over Time* (pp. 155-187). Emerald Publishing Limited.

Rosen, M. A., & Dietz, A. S. (2017). Team performance measurement. In E. Salas, R. Rico & J. Passmore, (Eds.), *The Wiley Blackwell Handbook of the Psychology of Team Working and Collaborative Processes* (pp. 479-502). Wiley Blackwell

Rosen, M. A., Salas, E., Wilson, K. A., King, H. B., Salisbury, M., Augenstein, J. S., ... & Birnbach, D. J. (2008). Measuring team performance in simulation-based training: adopting best practices for healthcare. *Simulation in Healthcare*, *3*(1), 33-41.

Rudolph, J. W., Simon, R., & Raemer, D. B. (2007). Which reality matters? Questions on the path to high engagement in healthcare simulation. *Simulation in Healthcare, 2*, 161-163.

Sambucini, V. (2017). Bayesian vs Frequentist Power Functions to Determine the Optimal Sample Size: Testing One Sample Binomial Proportion Using Exact Methods. In J.P. Tejedor (Eds.) *Bayesian Inference* (pp 77-97). InTech.

Salas, E., Shuffler, M. L., Thayer, A. L., Bedwell, W. L., & Lazzara, E. H. (2015). Understanding and improving teamwork in organizations: A scientifically based practical guide. *Human Resource Management*, *54*(4), 599-622.

Salas, E.M., Driskell, J.E., & Hughes, S. (1996). Introduction: The study of stress and human performance. In J.E. Driskell & E. Salas (Eds.), *Stress and Human Performance* (pp. 1-46). Hillsdale, NJ: Erlbaum.

Schmutz, J., & Manser, T. (2013). Do team processes really have an effect on clinical performance? A systematic literature review. *British Journal of Anaesthesia*, *110*(4), 529-544.

Schmutz, J. B., Lei, Z., Eppich, W. J., & Manser, T. (2018). Reflection in the heat of the moment: The role of in-action team reflexivity in health care emergency teams. *Journal of Organizational Behaviour*, *39*(6), 749-765.

Sexton, J. B., Thomas, E. J., & Helmreich, R. L. (2000). Error, stress, and teamwork in medicine and aviation: cross sectional surveys. *British Medical Journal, 320*(7237), 745-749.

Shuffler, M. L., & Carter, D. R. (2018). Teamwork situated in multiteam systems: Key lessons learned and future opportunities. *American Psychologist*, *73*(4), 390.

Sleeper, J. A., & Thompson, C. (2008). The use of hi fidelity simulation to enhance nursing students' therapeutic communication skills. *International Journal of Nursing Education Scholarship*, *5*(1), 1-12.

Sneddon, A., Mearns, K., & Flin, R. (2006). Situation awareness and safety in offshore drill crews. *Cognition, Technology & Work, 8*(4), 255-267

Stachowski, A. A., Kaplan, S. A., & Waller, M. J. (2009). The benefits of flexible team interaction during crises. *Journal of Applied Psychology*, *94*(6), 1536.

Stanton, N. A., & Roberts, A. P. J. (2018). Examining social, information, and task networks in submarine command and control. *IEEE Transactions on Human-Machine Systems*, *48*(3), 252-265.

Taylor, P. J., (June, 2013). How technology is revolutionizing our understanding of human cooperation (Inaugural lecture). Twente University Press. Downloaded on 20 June, 2019, from https://www.utwente.nl/en/academic-ceremonies/inaugural-lectures/booklets-inaugural-lectures/2007-2014/Oratieboekje_Taylor.pdf

Valentine, M. A., Nembhard, I. M., & Edmondson, A. C. (2015). Measuring teamwork in health care settings: a review of survey instruments. *Medical Care*, *53*(4), 16-30

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). Bayesian methods for advancing psychological science. *Psychonomic Bulletin and Review, 25* (1), 1-4

Vessey, W. B., & Landon, L. B. (2017). Team Performance in Extreme Environments. In In E. Salas, R. Rico & J. Passmore, (Eds.), *The Wiley Blackwell Handbook of the Psychology of Team Working and Collaborative Processes*, (pp. 531-553). Wiley Blackwell.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... & Matzke, D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35-57.

Walker, G. H., Gibson, H., Stanton, N. A., Baber, C., Salmon, P., & Green, D. (2006). Event analysis of systemic teamwork (EAST): a novel integration of ergonomics methods to analyse C4i activity. *Ergonomics*, *49*(12-13), 1345-1369.

Walker, G. H., Stanton, N. A., Baber, C., Wells, L., Gibson, H., Salmon, P., & Jenkins, D. (2010). From ethnography to the EAST method: A tractable approach for representing distributed cognition in Air Traffic Control. *Ergonomics*, *53*(2), 184-197.

Wallenius, C., Larsson, G., & Johansson, C. R. (2004). Military observers' reactions and performance when facing danger. *Military Psychology*, *16*(4), 211-229.

Waller, M. J., & Kaplan, S. A. (2018). Systematic behavioral observation for emergent team phenomena: Key considerations for quantitative video-based approaches. *Organizational Research Methods*, *21*(2), 500-515.

Wang, M., & Hanges, P. J. (2011). Latent class procedures: Applications to organizational research. *Organizational Research Methods*, *14*(1), 24-31.

Wauben, L. S. G. L., Dekker-van Doorn, C. M., Van Wijngaarden, J. D. H., Goossens, R. H. M., Huijsman, R., Klein, J., & Lange, J. F. (2011). Discrepant perceptions of communication, teamwork and situation awareness among surgical team members. *International Journal for Quality in Health Care*, *23*(2), 159-166.

Wetzel, C. M., Kneebone, R. L., Woloshynowych, M., Nestel, D., Moorthy, K., Kidd, J., & Darzi, A. (2006). The effects of stress on surgical performance. *The American Journal of Surgery*, *191*(1), 5-10.

Willemsen-Dunlap, A. M., Binstadt, E. S., Nguyen, M. C., Elliott, N. C., Cheney, A. R., Stevens, R. H., & Dooley-Hash, S. (2018). Alternative Markers of Performance in Simulation: Where We Are and Where We Need To Go. *Academic Emergency Medicine*, *25*(2), 250-254.

Wu, L., Waber, B., Aral, S., Brynjolfsson, E., & Pentland, A. (2008). Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an it configuration task. Retrieved from: SSRN:http://dx.doi.org/10.2139/ssrn.1130251

Young, D. L., Goodie, A. S., Hall, D. B., & Wu, E. (2012). Decision making under time pressure, modeled in a prospect theory framework. *Organizational Behavior and Human Decision Processes*, *118*(2), 179-188.

Yule, S., Flin, R., Maran, N., Rowley, D., Youngson, G., & Paterson-Brown, S. (2008). Surgeons' non-technical skills in the operating room: reliability testing of the NOTSS behavior rating system. *World Journal of Surgery*, *32*(4), 548-556.

Zaccaro, S. J., Gualtieri, J., & Minionis, D. (1995). Task cohesion as a facilitator of team decision making under temporal urgency. *Military Psychology*, *7*(2), 77-93.

Zhang, Y., Olenick, J., Chang, C. H., Kozlowski, S. W., & Hung, H. (2018). TeamSense: Assessing Personal Affect and Group Cohesion in Small Teams through Dyadic Interaction and Behavior Analysis with Wearable Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 150.

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management*, *41*(2), 390-420.