Volume 14 | Number 2 | June 2020

Journal *of* Cognitive Engineering *and* Decision Making
A Publication of the Human Factors and Ergonomics Society

CONTENTS

99 "You Can't Hide Your Lyin' Eyes": Investigating the Relationship Between Associative Learning, Cue Awareness, and Decision Performance in Detecting Lies
Ben W. Morrison, David Johnston, Mathew Naylor, Natalie M. V. Morrison, and Daniel Forrest

112 Rational Adaptation: Contextual Effects in Medical Decision Making
Frank Eric Robinson, Markus A. Feufel, Valerie L. Shalin, Debra Steele-Johnson, and Brian Springer

132 Effects of Workload and Workload Transitions on Attention Allocation in a Dual-Task Environment: Evidence From Eye Tracking Metrics
Nadine Marie Moacdieh, Shannon P. Devlin, Hussein Jundi, and Sara Lu Riggs

152 Improving Traffic Incident Management Using Team Cognitive Work Analysis
Vanessa Cattermole-Terzic and Tim Horberry

174 Transparency in Autonomous Teammates: Intention to Support as Teaming Information
April Rose Panganiban, Gerald Matthews, Michael D. Long

Journal *of* Cognitive Engineering *and* Decision Making

"You Can't Hide Your Lyin' Eyes": Investigating the Relationship Between Associative Learning, Cue Awareness, and Decision Performance in Detecting Lies

Ben W. Morrison, Australian College of Applied Psychology, Australia, and Charles Sturt University, Australia, **David Johnston**, Australian College of Applied Psychology, Australia, and Cambridge University, UK, **Mathew Naylor**, Australian College of Applied Psychology, Australia, and The University of Sydney, Australia, **Natalie M. V. Morrison**, Western Sydney University, Australia, and **Daniel Forrest**, The University of Sydney, Australia

Although skilled cue utilization is presumed to result primarily from domain-specific experience, individual differences in learning are theorized to play a significant role. Using a single-group correlational design, this study tested whether individuals' domain-general associative learning capacity was related to performance in a complex real-world decision task presumed to rely heavily on cues: lie detection. A total of 21 participants completed an associative learning task in the form of a Space Invaders-like game. In the game, those who learn the cues are able to respond faster to the appearance of an enemy ship. Participants were also surveyed on their awareness of cues in the game. This was followed by a lie detection task. It was hypothesized that greater associative learning would be associated with greater awareness of cues in the learning task, and subsequently, superior accuracy in the lie detection task. Participants' associative learning was correlated with their cue awareness (r_{pb} = .782, p < .001). Further, learning was associated with better performance in the lie detection task (r = .544, p = .011); however, accuracy was found to be unrelated to the types of cues reportedly used during detection. These findings have implications for our understanding of cue acquisition and expertise development.

Keywords: cue utilization, cue acquisition, expertise, lie detection, learning

Cues are events, whether internal (e.g., a mood) or external (e.g., a sound), that upon their recognition will signal significance to an individual. In this sense, cues "trigger" associations held in memory, which, when valid, enable us to make predictions (or diagnoses) about our environment (Wiggins, 2015). When engaged appropriately, cues enable relatively rapid and effective decision making (Klein, 2008; Mann et al., 2007). Cognitively demanding contexts (e.g., medicine, aviation, criminal investigation, clinical psychology, professional sport) appear to encourage individuals to engage cues regularly, given their ability to reduce cognitive load during decision making (Crane et al., 2018; Johnston & Morrison, 2016; Morrison et al., 2013; Morrison & Morrison, 2015; (Wiggins et al., 2014). Indeed, Easterbrook (1959) established that in such environments, which are typically characterized by an increased arousal state, decision makers' attention becomes more selective, targeting the most relevant cues for processing, while filtering out the superfluous ones

Greater cue utilization has been shown to be critical in the generation of efficient and accurate responses across a range of work domains (Brouwers et al., 2017; Gacasan & Wiggins, 2017; Morrison et al., 2018; Perry et al., 2013; Wiggins et al., 2018) and is a commonly cited strategy among experts (Johnston & Morrison, 2016; Kahneman & Klein, 2009; Loveday

Address correspondence to Ben W. Morrison, School of Psychology, Charles Sturt University, Building 1400, Panorama Avenue, Bathurst, NSW 2795, Australia, bmorrison@csu.edu.au.

Journal of Cognitive Engineering and Decision Making
2020, Volume 14, Number 2, June 2020, pp. 99–111
DOI: 10.1177/1555343420918084

et al., 2013). A process of cue refinement called cue discrimination has also been associated with expertise (Shanteau & Hall, 1992). Experts appear to rely on fewer cues than their novice counterparts, only selecting and using the most relevant ones (Shanteau, 1992). In a study that compared expert and novice critical care cardiovascular nurses, it was shown that only a limited number of cues were used by expert nurses across a range of decision scenarios (Reischman & Yarandi, 2002). These "critical" cues tend to be associated with fewer hypotheses, which assist in more timely and accurate assessments (i.e., they are relatively high in what has been termed "diagnosticity"; Schriver et al., 2008).

Many have highlighted the role of cue recognition in skilled intuition (Hertwig et al., 1999; Kahneman & Klein, 2009; Simon, 1992). In considering an expert's ability to rapidly size up a situation, Simon (1992) posits that "The situation has provided a cue: This cue has given the expert access to information stored in memory, and the information provides the answer. Intuition is nothing more and nothing less than recognition" (p. 155). The model of intuition as cue recognition has instigated a simple yet powerful explanatory framework in the naturalistic decision making (NDM) community.

The recognition-primed decision (RPD) model (Klein, 1993; Klein et al., 1986) extends on Simon's premise, offering an account of how experienced decision makers will leverage cues in the initial stages of situation assessment. The model posits that experienced decision makers will use cues to draw on a repertoire of stored patterns, which accumulate as a result of domain-specific experiences. In matching these patterns, the expert will nonconsciously generate and prioritize a series of plausible responses, which they can test via a process of mental simulation (a notion consistent with deGroot's concept of "progressive deepening"; 1978). Thus, for NDM researchers, cues are thought to initiate the intuitive aspect of the RPD model, which incorporates both intuitive and analytical phases. Indeed, in a dual-system model of cognition (Evans et al., 2009; Stanovich & West, 2000), cue utilization will occur as part of the associative machinery of System 1 (the rapid, automatic, unconscious, and relatively effortless processor), while simulation and deepening are better aligned to System 2 (slow, controlled, deliberate, and relatively effortful).

CUE ACQUISITION AND LEARNING

Cue utilization has been hypothesized to result from previous environmental experience (Wiggins, 2006). Decision makers are presumed to require an opportunity to observe potential cues and receive feedback upon their application (Kahneman & Klein, 2009). Here, it may be reasoned that deliberate engagement of System 2 during learning will work to sensitize our System 1 to the detection of relevant cues in the environment, harnessing the associative power of System 1. However, while experience with the operational environment is a necessary condition for the development of skilled intuition, it is not a sufficient one.

In addition to factors that influence our ability to acquire cues, such as the relative validity of the operational environment (Kahneman & Klein, 2009), it is apparent that some decision makers will learn more than others from equivalent experiences (Ericsson et al., 2006; Klein, 2015), and this will invariably influence their rate of cue acquisition. Indeed, a recent study questioned whether the capacity to learn cues may be partly trait-driven. Wiggins and Auton (2016) examined cue utilization among transmission power controllers and found that those who displayed greater cue utilization in the context of power control also displayed greater levels of cue utilization in a non-domain-specificcontext. This suggested that cue behaviors may be partly determined by inherent characteristics of the decision maker. Despite the consistent findings with regard to cue utilization and performance advantages, only limited research has explored the underlying mechanisms that would lead to individual differences in cue acquisition and utilization. This would appear to be theoretically relevant to the process of learning.

It can be reasoned that the initial identification and resultant utilization of cues must involve some learning process through which the association and meaning are established. Numerous theorists have developed explanatory

frameworks for the acquisition of associations in human memory (e.g., Anderson & Bower, 1973; Mandler, 2002). However, ultimately, the mechanism by which humans learn that two events are associated remains under contention.

A number of researchers have suggested that people learn through rule-based mechanisms involving the use of higher cognitive processes through a process of generation and evaluation about the associations (Lovibond & Shanks, 2002). This notion would imply significant involvement from System 2 mechanisms. Opposing this are researchers who maintain that learning can proceed as the result of capturing regularities within the environment, a stimulus-driven process whereby associations are automatically formed between these representations, favoring a System 1 view of development (Clark et al., 2002). Thus, one of the key differences between these theories of associative memory appears to be the degree of conscious awareness a learner has regarding stimulus contingencies.

While much of the NDM literature would support the notion that experts' tacit knowledge, including their use of cues, is typically unconscious and difficult to retrieve, cognitive interview techniques have proven successful in eliciting stimulus contingencies from experienced operators across numerous domains. For instance, in a recent study of cue utilization among Rugby League players, Johnston and Morrison (2016) used the Critical Decision Method (CDM; Crandall & Getchell-Reiter, 1993) to reveal that players' processes relating to cue utilization were consciously adaptive in response to a dynamic environment. As such, it was hypothesized that experts' awareness of cues allows for a continual reordering of cue significance. In contrast, novices are seemingly less aware of cues and hence, their utilization is less amenable to change.

Awareness has also been conceptualized as a powerful and sophisticated selection mechanism, which enables individuals to focus attention toward a restricted set of objects and events (Eriksen & Yeh, 1985; Treisman & Gelade, 1980). This is consistent with an information reduction hypothesis in the cognitive–perceptual skills literature (Haider & Frensch, 1999), which posits that much of the performance advantage associated with cue utilization stems from their attention management properties. Studies of awareness have thus concentrated on measures of attention.

The study of a phenomenon called "inattentional blindness," the failure to notice an object when attention is focused elsewhere (Mack & Rock, 1998), has revealed that regardless of whether the subjects noticed the unexpected object or not, all observers spent on average the same amount of time looking at the unexpected object (Koivisto et al., 2004). Further to this, in expert–novice domain-relevant paradigms of inattentional blindness, skilled performers were less susceptible to inattentional blindness in dynamic situations (Memmert, 2006). This suggests that awareness may also relate to cognitive efficiencies, a process modeled by Anderson's Adaptive Control of Thought—Rational (ACT-R) theory (1996). Despite the apparent usefulness of this model, no research has explored individual variations in learning and its subsequent effects on the development of these "chunks" and condition action statements.

Irrespective of the theoretical perspective one subscribes to regarding the mechanisms involved in the formation of associations, it is clear that their acquisition underpins learning. Thus, it is reasonable to posit that a domain-general associative learning capacity will likely play a fundamental role in the early stages of expertise development. This notion is comparable to how other cognitive abilities predict individual differences in complex task performance early in learning (Ward et al., 2019). However, it is typically less clear whether such abilities remain predictive after extensive task exposure.

Whilst it is presumed that such abilities are likely to be later mitigated or supplanted by domain-specific mechanisms, we posit that in many domains they will remain predictive nonetheless. Indeed, like other domain-general abilities (e.g., intelligence, attentional control, working memory capacity), associative learning capacity may partly explain why some individuals will advance to higher levels of expertise than others with similar experience levels. However, unlike other domain-general abilities, the relationship between learners'

domain-general associative learning capacity and decision-making performance in real-world contexts is largely untested. Further, investigations of the role of individuals' awareness of the associations in their learning, which would presumably have implications for instructional design (e.g., explicit instruction vs. guided discovery), are similarly limited. The current paper seeks to address these gaps by testing whether domain-general associative learning is associated with performance differences in a complex real-world task. The study also examined the degree of relation between learners' rate of domain-general associative learning and their awareness of acquired associations.

Lie Detection as a Context for Studying Cue Utilization

Lie detection represents a sound context for addressing our research questions as (1) it represents a complex real-world task in which objective performance is readily measured; (2) performance is presumed to rely heavily on the use of a broad range of cues (e.g., facial features, tone and pitch of voice, word/sentence length and complexity, speech rate, interaction behaviors with interviewers and co-conspirators); (3) it is a task that most people will have had extensive exposure to in their everyday lives (without formal training); and (4) some individuals are able to develop a degree of expertise in the domain, despite the lack of well-defined knowledge and rules.

Interest in lie detection has been a focus of study for almost half a century (Vrij, 2008). Researchers have systematically investigated the behaviors that probabilistically signal deception (DePaulo et al., 2003), and it is presumed that such behaviors are able to manifest in cue-based associations. Indeed, Ekman and Friesen (1969) theorized two broad categories of cues that emanate from the deceiver: leakage and deception cues. Leakage cues refer to nonverbal cues that reveal what the liar is trying to hide. Deception cues indicate that deception may be occurring without indicating the nature of the information that is being concealed.

Different approaches to nonverbal deceptive behavior have been theorized around the different control approaches used by the deceiver: the cognitive, emotional, and the attempted control approach (Vrij, 2008). The cognitive approach assumes that lying is more cognitively complex than telling the truth, hence liars will make more mistakes (Vrij, 2008). The emotional approach suggests that a liar can feel either guilt, fear, or excitement and that each emotional state will lead to an observable bodily motion such as microexpressions (Ekman et al., 1991). The attempted control approach proposes that a liar in their attempt to control specific aspects of their presentation will paradoxically display a cue to reveal deception that has occurred (Greene et al., 1985). Supporting these theories, in the most comprehensive analysis of cues to deception, DePaulo et al. (2003) showed that there existed a number of cues that may be helpful in distinguishing between lie and truth. Since then, other researchers have uncovered a range of additional cues (Driskell et al., 2012; Fuller et al., 2013; Vrij & Granhag, 2014).

Study Aims and Hypotheses

The aim of the current study was to test whether domain-general associative learning capacity was related to performance in a complex real-world decision task presumed to rely heavily on cues: lie detection. Further, we explored the relationship between decision makers' awareness of associations during learning and their domain-general associative learning capacity. Participants were required to complete an associative learning task in the form of a Space Invaders-like video game (consistent with the method of Forrest et al., 2016). In the game participants had to respond quickly to a target stimulus (an "enemy spaceship") that appeared intermittently among distractor stimuli ("friendly spaceships"). Although participants were given no indication that they could learn to predict the appearance of the enemy ship, specific distractor stimuli acted as a potential cue to its impending appearance. Thus, those who successfully learned the cue would be faster to respond to the appearance of the enemy ship. This task also included an explicit awareness measure developed by Forrest et al. (2016) to

gauge the participants' degree of awareness of the associations (i.e., the cues).

Participants then viewed 12 videos of people who were placed in a high-stakes mock crime paradigm (ascertained from ten Brinke et al., 2014). After each video, participants were required to answer a forced-choice question relating to whether or not the person in the video was lying or telling the truth.

If conscious awareness is a prerequisite to association acquisition, we would expect to see a positive relationship between participants' domain-general associative learning capacity and their awareness of stimulus contingencies in the associative learning task. We would also expect that those participants who demonstrate a greater capacity for domain-general associative learning will have developed a superior repertoire of domain-specific lie detection patterns over time, which will result in a concomitant advantage in detecting lies. Therefore, it was hypothesized that those participants demonstrating a greater degree of domain-general associative learning would show (1) greater awareness of the cues in the learning task and (2) greater accuracy in the lie detection task. Additionally, the cues identified by participants were categorized based on cue typology commonly cited in the extant literature (Sporer, 1997; Sporer & Schwandt, 2006): nonverbal visual (e.g., eye contact, head movements); paraverbal (e.g., response latencies, pauses); verbal content (e.g., use of certain words, consistency of a statement); combined verbal type (i.e., verbal and paraverbal); and combined (i.e., nonverbal visual, paraverbal, and verbal content). The categorization process enabled an investigation of the potential association between cue type and lie detection accuracy.

METHOD

Participants

Participants comprised a sample of convenience and were recruited through an online social media platform and a research recruitment portal (i.e., SONA) at the Australian College of Applied Psychology. There were 21 participants (18 female) aged between 23 and 47 (M = 30.57, SD = 6.93). All participants required normal color vision and hearing for eligibility. The research was approved by the institution's Human Research Ethics Committee.

Apparatus and Materials

The associative learning task acquired from Forrest et al. (2016) was programmed with Python on Windows 7 and was run on Macintosh Computers operating through Windows, connected to 22″ LCD monitors of 1920 × 1080 resolution, refreshed at a rate of 60 Hz. Responses were recorded using a standard computer keyboard.

Associative Learning Task. Learning task stimuli comprised nine differently colored and differently shaped images 200 × 200 pixels in size, each representing unique ships. The colors used were blue, cyan, green, light green, orange, pink, purple, red, and yellow. A 400 × 800 pixel image of a green and blue semicircle represented planet Earth at the bottom of the screen, and the outer border of the game area was a red semicircle. Distance from the red border to the edge of the planet measured 28 cm. All images appeared on a black background and the game-related text was white. See Figure 1 for a screen shot of the associative learning task.

Of the nine stimuli, eight were designated as "friendly" while one was an enemy ship. Six of the eight friendly ships were distractors. Purple was selected as the enemy ship after a previous experiment revealed it mostly effectively avoided the salience of brighter colors, had no preexisting stereotypes of meaning, and was differentiated from other colors more equally (Forrest et al., 2016). The other ships, both the cues (two) and distractors (i.e., no signal to enemy ship; six), were required to differ sufficiently in shape and color both to the enemy ship and to each other, and colors were again selected in line with Forrest et al.'s (2016) advice.

Accuracy in the learning task was measured as a percentage of the enemy ships that participants correctly eliminated before reaching the planet. The degree of associative learning was calculated by taking the difference between participants' response times (RTs in milliseconds; ms)

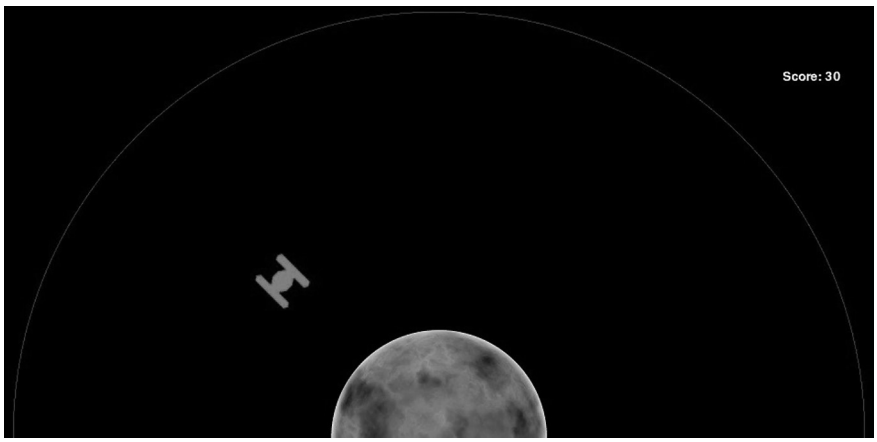


Figure 1. Example of the associative learning task game display.

in shooting the enemy ship in response to the cue ships, compared to shooting it in the absence of a prompt. In this sense, a greater difference in RT equated to greater evidence of domain-general associative learning.

Awareness Measure. Participants completed an explicit "awareness" measure (Forrest et al., 2016) to ascertain their understanding of the relationship between the cue ships and the enemy ship that they signaled. The questionnaire used a free recall component in which participants were asked whether they could predict the appearance of the enemy ship. Awareness was ascertained as to whether the participant could verbally identify the relationship between the enemy ship and the friendly ships that preceded it).

Lie Detection Task. Participants viewed 12 prerecorded interviews from a close midshot, whereby each character was visible from the middle chest upwards and the face clearly seen. The videos were ascertained from ten Brinke et al. (2014) who placed psychology students in a high-stakes mock crime paradigm (consistent with Kircher, Horowitz, & Raskin, 1988). In their study, participants (N = 12; 6 male, 6 female) were randomly assigned to steal or not steal $100 (steal condition, n = 6; no-steal condition, n = 6). Participants were told that they would earn the $100 if they convinced the experimenter that they had not stolen the money or lose $100 if they failed to convince the experimenter (regardless of whether they had). Participants were told that they would be entered into a lottery to win an additional $500 if they were successful. During the recorded interview they were asked a range of questions including baseline questions (e.g., "What are you wearing today?") and interview questions (e.g., "Did you steal the money?"). Each video lasted an average of 97 s (SD = 21.62 s).

Following the viewing of each video, participants provided a direct, self-report decision as to whether each character was lying (i.e., "In your opinion was the respondent lying?"). Next, participants were asked about the features of the video that informed their decision. Participants were provided an example to help elicit the potential cues that were used in the process of their decision making: "Consider you believe a person looked tired, the cues you may have used could be frequent yawning and red eyes." A lie detection accuracy score was computed for participants by summing the number of correct instances of lie detection across the 12 presentations.

Procedure

The participants completed the study in individual sessions. Participants were positioned to be oriented toward the screen and instructed that they would be required to maintain the same

position while playing the game. Participants were then instructed to place noise cancellation headphones over their ears, which would remain on for the remainder of the study. Participants were then instructed to play the computer game with the aim of scoring the highest amount of points possible.

The game involved colored spacecraft entering from two points around a semicircular boundary, 90° apart, and approaching a planet at the center in a direct trajectory. Text-based instructions then appeared on-screen informing participants to allow friendly ships to land and to prevent enemy ships from landing by pressing either the left or right shift keys to shoot ships from the left or right entry points, respectively. Doing so caused a red beam to emerge from the planet and destroy the ship on the corresponding side. For friendly ships, safe entrance to the planet was indicated by a short bubble animation, an accompanying positive sound, and a "10" pop-up. If destroyed, friendly ships showed a short neutral animation, a neutral sound, and a "−10" pop-up. When enemy ships were destroyed, an explosion animation and sound were played followed by a "50" pop-up whereas if not destroyed, a negative sound would play, the planet would turn red, and a "−50" pop-up was displayed. Participants were not told which ships were friendly or hostile. Instead, they were told that they would learn as they played.

The game consisted of three blocks, each lasting 306 seconds (s), allowing for two breaks. Stimuli included eight friendly ships and one enemy ship. Two of the friendly ships acted as potential cues, which reliably signaled the incoming enemy ship. Enemy ships reliably followed the cue ships, and occasionally followed a distractor stimulus. At the completion of one block, on-screen text informed participants to have a short break if required. Two further 306-s blocks followed, divided by a break, until the end of the third block where participants were informed that the study had ended and to wait for the researcher. Each trial lasted 1 s and only one stimulus was displayed on-screen at any given time. Each stimulus took 0.5-s to reach the planet. Thus, one trial consisted of 0.5-s stimulus exposure followed by a 0.5-s gap before the next trial began.

Following completion of this task, participants completed the awareness measure. They were required to fill out a form enquiring about what methods they used to predict the arrival of the oncoming enemy ship. Once completed the form was returned to the researcher.

The researcher then opened up an online questionnaire, which was displayed in Qualtrics (Qualtrics LLC, Provo, UT, USA). Participants were directed to follow instructions that appeared on-screen. Participants first completed a demographics questionnaire. Once completed participants were presented with an instruction informing them that "You are about to view a video of a person either lying or telling the truth. Pay close attention because you will be asked a series of questions about this." When ready, participants pressed the next button to continue and subsequently viewed the deception stimuli. Videos could only be viewed on a singular occasion. Following the viewing of the video, participants were then required to answer a series of questions relating to the potential presence of deception in the video and the features of the videos that they used to base their decision on. Once all questions were completed, participants would select the "next button" and be primed by an instruction screen. This screen informed the participant that they were about to view another video. The participant would select "next" before being able to view the next deception stimuli. Videos of lying and truth telling were randomly interspersed for the participants and followed the same trial structure of priming, stimuli, and questions about stimuli for a total of 12 trials. Following the completion of the 12 trials, participants were presented with an on-screen instruction informing them that they had completed the study. Participants were then debriefed by the researcher.

RESULTS

Lie Detection

Overall accuracy scores were computed through summing the lie detection accuracy in each of the 12 presentations, and ranged from 16.67% to 66.67% (M = 38.89, SD = 16.94, N = 21).

Associative Learning, Awareness, and Lie Detection Accuracy

Mean reaction times (RTs) were recorded in ms, and the degree of associative learning was calculated by taking the difference between participants' RT in shooting the enemy ship in response to the cue ships, compared to shooting it in the absence of a prompt. In this sense, a greater difference in RT equated to greater evidence of domain-general associative learning.

Individuals were coded based on their "awareness." Awareness was ascertained as to whether the participant could verbally identify the relationship between the enemy ship and the friendly ships that preceded it. After meeting the associated mathematical assumptions, data were subjected to correlational analyses to investigate the potential relationship between domain-general associative learning capacity, awareness of cues in the learning task, and subsequent accuracy in the lie detection task. A point-biserial correlational analysis revealed a statistically significant positive correlation between associative learning and awareness, r_{pb} $(N = 21) = .782$, $p < .001$ (large effect). Similarly, Pearson's correlational analysis revealed a statistically significant positive correlation between associative learning and lie detection performance, $r\ (N = 21) = .544$, $p = .011$ (large effect). These results indicate that a greater degree of associative learning in the initial learning task was associated with greater awareness of cues present in the learning task and greater performance on the lie detection task.

Deception Cues

A directed content analysis was conducted to categorize the cues that were used by each participant in the analysis. Cue categorization was based on cue typology commonly cited in the extant literature (Sporer, 1997; Sporer & Schwandt, 2006): nonverbal visual (e.g., eye contact, head movements), paraverbal (e.g., response latencies, pauses), and verbal content cues (e.g., use of certain words, consistency of a statement), combined verbal type (i.e., verbal and paraverbal), and combined (i.e., nonverbal visual, paraverbal, and verbal content cues).

The categorization process enabled an investigation of the potential association between cue type and lie detection accuracy; however, a two-way χ^2 revealed a nonsignificant association, $\chi^2 (4, N = 21) = 2.373$, $p = .667$. The frequencies are shown in Table 1.

DISCUSSION

The aim of the current study was to test whether domain-general associative learning capacity was related to performance in a complex real-world decision task presumed to rely heavily on cues: lie detection. Further, we explored the relationship between decision makers' awareness of associations and their associative learning capacity. As hypothesized, the results revealed a significant positive relationship between participants' domain-general associative learning capacity and their awareness of cues in the learning task. Further, learning was found to be positively associated with

TABLE 1: Frequencies, Percentages, and Adjusted Standardized Residuals (ASR) for Correct and Incorrect Lie Detection When Using the Five Different Cue Types

	Correct Lie Detection				Incorrect Lie Detection				
Cue Type	*f*	%	f_e	ASR[a]	*f*	%	f_e	ASR[a]	Total
Nonverbal visual	22	45.8	20.5	0.5	26	54.2	27.5	−0.5	48
Paraverbal	8	32	10.7	−1.2	17	68	14.3	1.2	25
Verbal content	13	38.2	14.6	−0.6	21	61.8	19.4	0.6	34
Combined	51	44	49.6	0.4	65	56	66.4	−0.4	116
Verbal combined	7	53.8	5.6	0.8	6	46.2	7.4	−0.8	13

lie detection performance. Additionally, the cues identified by participants were categorized based on cue typology commonly cited in the extant literature (Sporer, 1997; Sporer & Schwandt, 2006); however, no association was found between cue type and lie detection accuracy.

Lie Detection

The current findings regarding lie detection accuracy rates (38.89%) were of dissimilar magnitude to those reported in a previous meta-analysis (54%; Bond & DePaulo, 2006). This difference could be a reflection in the variability of stimuli used across studies. Indeed, the current findings were more similar to ten Brinke et al. (2014) who, using the method adopted in the current study, found a mean detection accuracy rate of 43.75%.

Associative Learning

As anticipated, "aware" participants demonstrated a superior associative learning rate compared to those who were found to be "unaware." Forrest et al. (2016) explained that this was due to participants having conscious awareness of contingencies, which was priming faster reaction times to cued trials. Unexpectedly, compared to Forrest et al. (2016), proportionately fewer participants reported an explicit awareness of the cue ships that signaled the enemy ship, only 14% compared to 25% in Forrest et al. (2016). This may be due to adaptations in the task, which involved participants' active involvement (pressing space in response to stimuli) as opposed to Forrest et al.'s (2016) design, which involved a noninstructed observation. This could have promoted the use of more complex strategies rather than a search for simple associations.

Deception Cues

The analysis revealed that there was no one category of cue that was better able to identify deception. This would support a previous meta-analysis that revealed that there exists no one consistent and reliable indicator to detecting deception (Bond & DePaulo, 2006). Within the context of this study, participants reported that they tended to use multiple (i.e., combined) cues in trying to detect deception; however, this strategy was also not found to be more accurate than any singular cue. These findings may also allude to participants' difficulties in consciously recognizing and/or articulating the cues they use. This is consistent with the findings of ten Brinke et al. (2014) who found that indirect, nonconscious measures of deception detection were significantly more accurate than direct, explicit measures.

Our findings regarding the types of cues used may also have been stifled by the artificial nature of the task and the static nature of the available cues. Indeed, recent trends in lie detection findings emphasize the importance of the active interaction between the observer and the deceiver in the inducement of deception cues (e.g., Hartwig et al., 2014). Future studies examining cue utilization in this context should consider the design of methods that enable a more dynamic interaction between observer and deceiver. Further, greater attention should be paid to cultural differences that may greatly influence cue meaning (e.g., pitch of voice has been shown to signal deception in some cultures, and conversely, truth-telling in others; Matsumoto et al., 2015).

Associative Learning, Awareness, and Lie Detection Accuracy

In line with the researchers' hypothesis, there existed a significant relationship between participants' domain-general associative learning capacity and lie detection accuracy. Previous research has highlighted that individual differences with regard to learning exist; however, the implications of these differences have not been explored. The finding that an individual's domain-general associative learning capacity correlated with performance might partly explain why the ability to detect deceit has been found to generalize across different scenarios (Frank & Ekman, 1997).

The current findings are consistent with the ACT-R framework (Anderson, 1996). The ACT-R theory provides an understanding of how knowledge structures are formed and activated in the performance of a skill. ACT-R

proposes that learning is accomplished in two stages. First, the user must learn the facts and store the facts as chunks in memory. The first stage has two components, a learning process and a storage phase. Second, these chunks of facts must be converted into decision rules. This study investigated one component of this first stage (learning) and showed that this process had cumulative effects on the whole system. It is hypothesized that awareness is related more significantly to the second stage where what is learned is reformulated into decision rules. It is suggested that the interplay between a learning system and awareness of associations may represent qualities leading to enhanced cognitive development. This could have implications for training, as it would suggest that interventions could be tailored for either those who require exposure-based training (i.e., "unaware") or those who require more explicit instructions (i.e., "aware"). However, we underline here that while our findings are superficially consistent with an ACT-R framework, the methods we have employed do not provide a direct test of cognitive architecture. As such, our conclusions are constrained to the identification of a domain-general associative learning mechanism that may play a role in to-be-acquired domain-specific mechanisms. More work is required to gain understanding in how the domain-general mechanisms identified here contribute to valid cue utilization behaviors, which will primarily arise out of extensive domain-specific experiences. The outcomes of such work would undoubtedly have significant theoretical and applied implications. For instance, understanding the extent to which simple, domain-general measures remain predictive of complex task performance in specific domains has the potential to inform training and selection procedures.

Limitations

The approach taken in this research contributes to theoretical models of expertise development but carries with it limitations. One significant issue stems from the fact that the same outcome can be produced by indiscernible variations between higher level cognitive structures and lower level processes. This is related to the context-dependency issue, reflected in the problem of one–many relations (Zalta, 2005). It is known that there is a one–many relation between neural pathways and higher level cognitions, such that a learning mechanism can causally lead to or be part of different higher level states depending on the context in which it is activated (Smith & Vela, 2001).

It must be noted that while "aware" participants in our study appeared to have an advantage in this domain, the factors that mediate the space between domain-general learning abilities and domain-specific decision performance have not been explored here. Indeed, while it is argued here that formation of associations represent the most fundamental building block of skilled intuition, it is presumed that much of the advantage to decision performance will arise from a range of cognitive skills not examined here (e.g., problem detection, sensemaking, and uncertainty management). Thus, it is with caution that the current findings should be interpreted.

The approach to studying the nature of cue utilization adopted in the study was also somewhat limited. A range of methodologies have been developed for investigating cue-related behavior, such as eye-tracking techniques, response latency-based recognition tasks (Morrison et al., 2013), and cue assessment batteries (e.g., EXPERTise; Loveday et al., 2014). These methods would provide a richer and more precise set of data in relation to users' cue search behavior, recognition, and discrimination and should be considered in extending the current line of investigation. Further, those wishing to continue using the lie detection paradigm as a context for studying expertise development may consider other known factors that may influence cue utilization in the domain (e.g., the use of global heuristics and truth biases; Feeley & deTurck, 1995).

Finally, while the sample size employed in the study is considered adequate for a single-group design (Field, 2013), it may somewhat limit the strength of the conclusions drawn here. The behavioral nature of the methodology employed and the time commitment required from participants presents challenges for future researchers wishing to extend on the current design.

Future Directions

Whilst this research has shown a correlation between learning differences and performance in a complex real-world task, it demands a deeper investigation into these underlying processes. As an example, individual differences in perception of time have been observed (Gilaie-Dotan et al., 2010), but no study has focused on the broader implications. Of specific interest to deception, microexpressions occur at a time threshold of 230 ms, which falls outside levels of conscious awareness (Yan et al., 2013). This would suggest that time perception would invariably play a role in having the capacity to observe and a learning process efficient enough to analyze this. However, the debate within psychology surrounding unconscious processes influencing the conscious is controversial. Whilst neuroscience appears to support the notions of an automatic associative system (Bargh & Morsella, 2008), the learning field continues to advocate for the causal role of awareness in learning (Weidemann et al., 2016).

The role of feedback in cue acquisition also requires closer investigation. The learning of "real-world" cues will require a degree of feedback to reinforce the fledgling association, which may be less available or timely than the feedback seen in the associative learning task used here (Kahneman & Klein, 2009). How decision makers learn cues in contexts with nonexplicit or no feedback is largely unclear. This is potentially problematic for work domains where formulated decisions are met with delayed, minimal, or even no feedback regarding their efficacy. For instance, a health practitioner may receive only sporadic and delayed feedback about a diagnosis or choice of intervention. In such cases, the decision maker is presumably limited in their capacity to learn about potential predictive cues that might help develop their expertise.

One current hypothesis is that learning can be driven by decision makers' confidence when performance feedback is of poor quality or absent. Hainguerlot et al. (2018) demonstrated that successful learning of the predictive value of cues in the absence of external feedback was positively related to decision makers' confidence judgments, which was understood to be a superior meta-cognitive ability to distinguish between correct responses and errors. The role of such metacognitive abilities in the acquisition of cue-based associations warrants further investigation, particularly in domains characterized by a limited opportunity for explicit feedback.

Conclusion

The current findings suggest that participants' inherent capacity for domain-general associative learning can be related to their decision performance in complex real-world decision tasks, in this case, lie detection. Further, this learning capacity was found to be positively related to participants' conscious awareness of learned associations. While the factors that mediate the space between the domain-general learning abilities and domain-specific decision performance have not been explored here, the findings offer insight into the initial processes involved in cue acquisition and expertise development.

ORCID iD

Ben W. Morrison https://orcid.org/0000-0002-5026-4675

REFERENCES

Anderson, J. R. (1996). Act: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355–365. https://doi.org/10.1037/0003-066X.51.4.355

Anderson, J. R., & Bower, G. (1973). *Human associative memory*. distributed by the Halsted Press division of John Wiley & Sons, Washington, D.C.

Bargh, J. A., & Morsella, E. (2008). The unconscious mind. *Perspectives on Psychological Science*, *3*(1), 73–79. https://doi.org/10.1111/j.1745-6916.2008.00064.x

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234. https://doi.org/10.1207/s15327957pspr1003_2

Brouwers, S., Wiggins, M. W., Griffin, B., Helton, W. S., & O'Hare, D. (2017). The role of cue utilisation in reducing the workload in a train control task. *Ergonomics*, *60*(11), 1500–1515. https://doi.org/10.1080/00140139.2017.1330494

Clark, R. E., Manns, J. R., & Squire, L. R. (2002). Classical conditioning, awareness, and brain systems. *Trends in Cognitive Sciences*, *6*(12), 524–531. https://doi.org/10.1016/S1364-6613(02)02041-7

Crandall, B., & Getchell-Reiter, K. (1993). Critical decision method: A technique for eliciting concrete assessment indicators from the "intuition" of NICU nurses. *Advances in Nursing Sciences*, *16*(1), 42–51.

Crane, M. F., Brouwers, S., Wiggins, M. W., Loveday, T., Forrest, K., Tan, S. G. M., & Cyna, A. M. (2018). "Experience isn't everything": How emotion affects the relationship between experience and cue utilization. *Human Factors: The Journal of*

the Human Factors and Ergonomics Society, *60*(5), 685–698. https://doi.org/10.1177/0018720818765800

deGroot, A. D. (1978). *Thought and choice in chess*. Mouton.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74–118.

Driskell, J. E., Salas, E., & Driskell, T. (2012). Social indicators of deception. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(4), 577–588. https://doi.org/10.1177/0018720812446338

Easterbrook, J. A. (1959). The effect of emotion on cue utilization and the organization of behavior. *Psychological Review*, *66*(3), 183–201. https://doi.org/10.1037/h0047707

Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry*, *32*(1), 88–106. https://doi.org/10.1080/00332747.1969.11023575

Ekman, P., O'Sullivan, M., Friesen, W. V., & Scherer, K. R. (1991). Invited article: Face, voice, and body in detecting deceit. *Journal of Nonverbal Behavior*, *15*(2), 125–135. https://doi.org/10.1007/BF00998267

Ericsson, K. A., Charness, N., Hoffman, R. R., & Feltovich, P. J. (Eds.). (2006). *The Cambridge handbook of expertise and expert performance*. Cambridge University Press.

Eriksen, C. W., & Yeh, Y. Y. (1985). Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Perception and Performance*, *11*(5), 583. https://doi.org/10.1037//0096-1523.11.5.583

Evans, J., St B. T., & Frankish, K. (Eds.). (2009). *In two minds: Dual processes and beyond*. Oxford University Press.

Feeley, T. H., & deTurck, M. A. (1995). Global cue usage in behavioral lie detection. *Communication Quarterly*, *43*(4), 420–430. https://doi.org/10.1080/01463379509369989

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th Ed). SAGE Publications Inc.

Forrest, D. R. L., Mather, M., & Harris, J. A. (2016). Unmasking latent inhibition in humans. *The Quarterly Journal of Experimental Psychology*, 1–18.

Frank, M. G., & Ekman, P. (1997). The ability to detect deceit generalizes across different types of high-stake lies. *Journal of Personality and Social Psychology*, *72*(6), 1429–1439. https://doi.org/10.1037/0022-3514.72.6.1429

Fuller, C. M., Biros, D. P., Burgoon, J., & Nunamaker, J. (2013). An examination and validation of linguistic constructs for studying high-stakes deception. *Group Decision and Negotiation*, *22*(1), 117–134. https://doi.org/10.1007/s10726-012-9300-z

Gacasan, E. M. P., & Wiggins, M. W. (2017). Sensemaking through cue utilisation in disaster recovery project management. *International Journal of Project Management*, *35*(5), 818–826. https://doi.org/10.1016/j.ijproman.2016.09.009

Gilaie-Dotan, S., Kanai, R., & Rees, G. (2010). Individual differences in time perception indicate different modality-independent mechanisms for different temporal durations. *Journal of Vision*, *10*(7), 1407–1407. https://doi.org/10.1167/10.7.1407

Greene, J. O., O'Hair, H., Cody, M. J., & Yen, C. (1985). Planning and control of behavior during deception. *Human Communication Research*, *11*(3), 335–364. https://doi.org/10.1111/j.1468-2958.1985.tb00051.x

Haider, H., & Frensch, P. A. (1999). Information reduction during skill acquisition: The influence of task instruction. *Journal of Experimental Psychology: Applied*, *5*(2), 129–151.

Hainguerlot, M., Vergnaud, J.-C., & de Gardelle, V. (2018). Metacognitive ability predicts learning cue-stimulus associations in the absence of external feedback. *Scientific Reports*, *8*(1), 5602. https://doi.org/10.1038/s41598-018-23936-9

Hartwig, M., Granhag, P. A., & Luke, T. (2014). Strategic use of evidence during investigative interviews: The state of the science. In D. C. Raskin, C. R. Honts, & J. C., Kircher (Eds.), *Credibility assessment: Scientific research and applications* (pp. 1–36). Academic Press.

Hertwig, R., Hoffrage, U., & Martingnon, L. (1999). Quick estimation: Letting the environment do the work. In G. Gigerenzer, P. M. Todd, & A. B. C. Research Group (Eds.), *Simple heuristics that make us smart* (pp. 37–58). Oxford University Press.

Johnston, D., & Morrison, B. W. (2016). The application of naturalistic decision-making techniques to explore cue use in rugby League Playmakers. *Journal of Cognitive Engineering and Decision Making*, *10*(4), 391–410. https://doi.org/10.1177/1555343416662181

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515, 526–526. https://doi.org/10.1037/a0016755

Kircher, J. C., Horowitz, S. W., & Raskin, D. C. (1988). Meta-analysis of mock crime studies of the control question polygraph technique. *Law and Human Behavior*, *12*(1), 79–90.

Klein, G. (2008). Naturalistic decision making. *Human Factors*, *50*(3), 456–460. https://doi.org/10.1518/001872008X288385

Klein, G. A. (1993). A recognition-primed decision (RPD) model of rapid decision making. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 138–147). Ablex.

Klein, G. (2015). A naturalistic decision making perspective on studying intuitive decision making. *Journal of Applied Research in Memory and Cognition*, *4*(3), 164–168.

Klein, G. A., Calderwood, R., & Clinton-Cirocco, A. (1986). Rapid decision making on the fireground. In *Proceedings of the Human Factors and Ergonomics Society 30th Annual Meeting* (Vol. 1, pp. 576–580). Ablex.

Koivisto, M., Hyönä, J., & Revonsuo, A. (2004). The effects of eye movements, spatial attention, and stimulus features on inattentional blindness. *Vision Research*, *44*(27), 3211–3221. https://doi.org/10.1016/j.visres.2004.07.026

Loveday, T., Wiggins, M. W., Harris, J. M., O'Hare, D., & Smith, N. (2013). An objective approach to identifying diagnostic expertise among power system controllers. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *55*(1), 90–107. https://doi.org/10.1177/0018720812450911

Loveday, T., Wiggins, M. W., & Searle, B. J. (2014). Cue utilization and broad indicators of workplace expertise. *Journal of Cognitive Engineering and Decision Making*, *8*(1), 98–113. https://doi.org/10.1177/1555343413497019

Lovibond, P. F., & Shanks, D. R. (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*(1), 3–26.

Mack, A., & Rock, I. (1998). *Inattentional blindness* (Vol. 33). MIT press.

Mandler, G. (2002). Origins of the cognitive (r)evolution. *Journal of the History of the Behavioral Sciences*, *38*(4), 339–353. https://doi.org/10.1002/jhbs.10066

Mann, D. T. Y., Williams, A. M., Ward, P., & Janelle, C. M. (2007). Perceptual-cognitive expertise in sport: A meta-analysis. *Journal of Sport and Exercise Psychology*, *29*(4), 457–478. https://doi.org/10.1123/jsep.29.4.457

Matsumoto, D., Hwang, H. C., & Sandoval, V. A. (2015). Ethnic similarities and differences in linguistic indicators of veracity and lying in a moderately high stakes scenario. *Journal of Police and Criminal Psychology*, *30*(1), 15–26. https://doi.org/10.1007/s11896-013-9137-7

Memmert, D. (2006). The effects of eye movements, age, and expertise on inattentional blindness. *Consciousness and Cognition*, *15*(3), 620–627. https://doi.org/10.1016/j.concog.2006.01.001

Morrison, B. W., & Morrison, N. M. V. (2015). Diagnostic cues in major crime scene investigation. In M. W. Wiggins & T. Loveday (Eds.), *Diagnostic expertise in organisational environments* (pp. 91–98). Ashgate.

Morrison, B. W., Morrison, N. M. V., Morton, J., & Harris, J. (2013). Using critical-cue inventories to advance virtual patient technologies in psychological assessment. In H. Shen, R. Smith, J. Paay, P. Calder, & T. Wyeld (Eds.), *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration (OzCHI '13)* (pp. 531–534). ACM.

Morrison, B. W., Wiggins, M. W., Bond, N. W., & Tyler, M. D. (2013). Measuring relative cue strength as a means of validating an inventory of expert offender profiling cues. *Journal of Cognitive Engineering and Decision Making*, *7*(2), 211–226. https://doi.org/10.1177/1555343412459192

Morrison, B. W., Wiggins, M. W., & Morrison, N. M. V. (2018). Utility of expert cue exposure as a mechanism to improve decision-making performance among novice criminal investigators. *Journal of Cognitive Engineering and Decision Making*, *12*(2), 99–111. https://doi.org/10.1177/1555343417746570

Perry, N. C., Wiggins, M. W., Childs, M., & Fogarty, G. (2013). The application of reduced-processing decision support systems to facilitate the acquisition of decision-making skills. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *55*(3), 535–544. https://doi.org/10.1177/0018720812467367

Reischman, R. R., & Yarandi, H. N. (2002). Critical care cardiovascular nurse expert and novice diagnostic cue utilization. *Journal of Advanced Nursing*, *39*(1), 24–34. https://doi.org/10.1046/j.1365-2648.2000.02239.x

Schriver, A. T., Morrow, D. G., Wickens, C. D., & Talleur, D. A. (2008). Expertise differences in attentional strategies related to pilot decision making. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(6), 864–878. https://doi.org/10.1518/001872008X374974

Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes*, *53*(2), 252–266. https://doi.org/10.1016/0749-5978(92)90064-E

Shanteau, J., & Hall, B. (1992). How much information does an expert use? Is it relevant? *Acta Psychologica*, *81*(1), 75–86. https://doi.org/10.1016/0001-6918(92)90012-3

Simon, H. A. (1992). What is an "Explanation" of behavior? *Psychological Science*, *3*(3), 150–161. https://doi.org/10.1111/j.1467-9280.1992.tb00017.x

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review*, *8*(2), 203–220. https://doi.org/10.3758/BF03196157

Sporer, S. L. (1997). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experienced events. *Applied Cognitive Psychology*, *11*(5), 373–397.

Sporer, S. L., & Schwandt, B. (2006). Paraverbal indicators of deception: A meta-analytic synthesis. *Applied Cognitive Psychology*, *20*(4), 421–446.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435

ten Brinke, L., Stimson, D., & Carney, D. R. (2014). Some evidence for unconscious lie detection. *Psychological Science*, *25*(5), 1098–1105. https://doi.org/10.1177/0956797614524421

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. Wiley.

Vrij, A., & Granhag, P. A. (2014). Eliciting information and detecting lies in intelligence interviewing: An overview of recent research. *Applied Cognitive Psychology*, *28*(6), 936–944. https://doi.org/10.1002/acp.3071

Ward, P., Maarten Schraagen, J., Gore, J., Roth, E., Hambrick, D., Burgoyne, A., & Oswald, F. (2019). Domain-General Models of Expertise: The Role of Cognitive Ability. In *The Oxford Handbook of expertise*. Oxford University Press.

Weidemann, G., Satkunarajah, M., & Lovibond, P. F. (2016). I think, therefore eyeblink: The importance of contingency awareness in conditioning. *Psychological Science*, *27*(4), 467–475.

Wiggins, M. W., Azar, D., Hawken, J., Loveday, T., & Newman, D. (2014). Cue-utilisation typologies and pilots' pre-flight and in-flight weather decision-making. *Safety Science*, *65*, 118–124.

Wiggins, M. W. (2006). Cue-based processing and human performance. In W. Karwowski (Ed.), *International encyclopedia of Ergonomics and human factors* (2nd ed.pp. 641–645). Taylor & Francis.

Wiggins, M., & Auton, J. (2016). *Trait-based cue utilisation in diagnostic settings*. Paper presented at the 2016 APS Congress, Melbourne, Victoria, Australia.

Wiggins, M. W. (2015). Cues in diagnostic reasoning. In M. W. Wiggins & T. Loveday (Eds.), *Diagnostic Expertise in Organizational Environments* (pp. 1–11). Ashgate Publishing.

Wiggins, M. W., Whincup, E., & Auton, J. C. (2018). Cue utilisation reduces effort but increases arousal during a process control task. *Applied Ergonomics*, *69*, 120–127. https://doi.org/10.1016/j.apergo.2018.01.012

Yan, W. J., Wu, Q., Liang, J., Chen, Y.-H., & Fu, X. (2013). How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of Nonverbal Behavior*, *37*(4), 217–230. https://doi.org/10.1007/s10919-013-0159-8

Zalta, E. (2005). *Stanford encyclopedia of philosophy*. Stanford University.

Ben W. Morrison is a senior lecturer and organizational psychologist working at Charles Sturt University, Australia. His research focuses on psychology in the workplace, including areas relating to human factors, cognitive engineering, cybersecurity, safety, human–computer interaction, automation, and expertise development.

David Johnston is a registered psychologist with research experience in clinical and sporting domains. He has experience working clinically in inpatient and community settings specializing in mood and anxiety disorders. David is an award winning researcher, winning the Australian Psychological Society Prize for his research investigating expert decision making in sport.

Mathew Naylor is a PhD candidate who has completed a Bachelor of Psychological Sciences (Honors) with a first class award. He is currently investigating how virtual reality can be combined with mindfulness and relaxation in order to help office workers to better cope with stress. Matt has an interest in cyberpsychology, video games, and gaming and internet cultures.

Natalie M. V. Morrison (PhD, MPsychClin) is a registered psychologist who works in both acute and community mental health settings. She presently works as an academic at the School of Medicine, Western Sydney University, teaching mental health to undergraduate and postgraduate medical students. Natalie's research interests started in cognitive psychology, specifically attentional modeling, but has since extended into areas related to trauma, chronic pain, and suicide prevention.

Daniel Forrest is a provisionally registered psychologist with experience in experimental and clinical research. His research areas include human and animal learning, memory, pain, and mindfulness-based interventions.

Rational Adaptation: Contextual Effects in Medical Decision Making

Frank Eric Robinson, Wright State University, Dayton, OH, USA, **Markus A. Feufel**, Technische Universität Berlin, Germany, **Valerie L. Shalin**, **Debra Steele-Johnson**, and **Brian Springer**, Wright State University, Dayton, OH, USA

Research and practice in medical decision making value consistency with standardized intervention, potentially neglecting the impact of various environmental features such as workload or the constraints of local work practice. This study presents both qualitative and quantitative analyses of emergency physicians' decision-making processes in their natural work setting to examine the impact of contextual features. We study contextual effects on two separable decision-making processes identified in quantified observational data: goal enactment and goal establishment. Whereas goal enactment responds to hospital differences and patient difficulty as main effects, goal establishment responds to their interaction. Our emphasis on goal establishment expands the scope of a medical decision-making literature focused on diagnosis, and extends to other professions and the more general conceptualization of expertise. From a theoretical perspective, we emphasize the importance of accounting for contextual variability within the bounds of expert behavior. Practically, we provide real-world examples of context effects that bear on the standardization of care, cost differences between hospitals, and the conceptualization of quality medical care.

Keywords: medical decision making, naturalistic observation, situated cognition, emergency medicine, expert reasoning

Address correspondence to Frank Eric Robinson, Department of Psychology, Wright State University, 2624 Q St., Bldg 851, Area B, Wright-Patterson AFB, OH 45433, USA, frank.robinson.5@us.af.mil.

Journal of Cognitive Engineering and Decision Making
2020, Volume 14, Number 2, June 2020, pp. 112–131
DOI: 10.1177/1555343420903212

INTRODUCTION

Both medically oriented and psychological research on medical decision making contribute to an impression of context independence. Departures from a consistent, prescriptive approach to decision making are often characterized as unwanted and are of particular concern in a regulated and litigious culture. However, some features of care delivery, including social consequences (e.g., Singh, 2018), available resources, or the constraints of a broader work system, properly influence physicians' decision making (Feufel, 2018; Greenhalgh et al., 2014; Timmermans & Mauck, 2005). The numerous subsystems of the decentralized U.S. healthcare system such as hospitals, pharmacies, clinics, and labs, with their own unique goals, values, beliefs, and norms (Carayon et al., 2012), provide multiple contexts that are certain to influence individuals' decision making.

Emergency medicine provides an opportunity to observe many of the above contextual features. Emergency physicians' core role is to determine the nature and severity of a patient's complaint, and whether and where in the local healthcare system the patient can access appropriate treatment (Feufel, 2009). Emergency physicians must also deal with nonmedical issues and simultaneously coordinate with multiple other specialties and medical facilities to secure follow-up care (Feufel, 2009).

We investigate the influence of context on the decision making that occurs in U.S. emergency medicine departments (EDs). We gather a sufficient number of observations to employ quantitative analysis, which guides the presentation of qualitative illustrations. Rather than focus on the decision itself, which varies across episodes, we focus on the naturally verbalized *processes* that

lead to a decision. Using quantitative multilevel modeling, we examine the effect of hospital setting, patient difficulty, and shift difficulty on two phases of the decision-making process that emerged from a factor analysis: goal establishment and goal enactment. Goal establishment is highly context sensitive, reflecting the interaction of patient difficulty and hospital setting. In contrast, goal enactment is less discretionary, falling under the accepted work practice of medicine, but modified by the available, often hospital-specific resources. Goal enactment responds to hospital as a main effect. Below we first provide an overview of the existing literature on the study of reasoning, particularly medical reasoning. Next, we summarize the logic of multilevel modeling, which allows us to quantify the effects of context, such as hospital, on medical decision-making behavior and separately examine the effects of patient difficulty, shift difficulty, and hospital, while statistically accounting for physician experience.

Studies of Medical Decision Making

The controlled experiment dominates contemporary science, wherein variables that are not of interest are controlled to establish the causal effects of a manipulated variable on an outcome. We see the effect of this approach in two complementary dimensions of empirical work on decision making: the definition of tasks (and independent variables), and the related issue of performance measurement (i.e., dependent variables). Such studies are not intended to address the contextual effects of interest here.

The definition of decision-making tasks. Medical decision-making studies traditionally examine a single patient's clinically relevant symptoms and related diagnoses. Researchers seek to capture physicians' reasoning as they reach a diagnosis or disposition, which is already known to the experimenters (e.g., Reyna & Lloyd, 2006). The focus on diagnosis as the crux of medical reasoning has persisted (Elstein, 2009), but has grown to address novel, hitherto unseen cases in separating true experts from the merely competent (Ward et al., 2018) and the influence of clinical context and variation (for a study of decision practices in general physicians, see Donner-Banzhoff et al., 2017). However, the varied instantiations of even basic health complaints and the influence of context on decision making cast doubt on whether "routine" problems truly exist in domains such as medicine. Doctors are far more likely to encounter many variants of the same problem in different contexts (e.g., treating different presentations of cardiac ischemia in two different hospitals) than a truly novel problem such as a rare disease. Thus we must account for both sources of novelty when evaluating decision making.

The focus on diagnosis prematurely eliminates numerous other tasks that are potentially associated with clinical reasoning (Shalin & Bertram, 1996; Shalin et al., 1997). For instance, Lippa et al. (2017) showed that clinical decisions in both acute and chronic care settings may vary based on patient input, preference, ability for shared decision making, and the availability of resources or support from family or other care providers. Consistent with Feufel (2009), Klein (2007a, 2007b), and Shalin and Bertram (1996), we suggest that doctors must dynamically determine their patients' problems and care goals over time, rather than solve predetermined, delimited, static, or rare diagnoses. The traditional experimental paradigm for studying medical decisions fails to capture processes related to goal establishment, the effects of context on these processes, and the iterative, hypothesis-driven nature of information gathering in the real world (Kulatunga-Moruzi et al., 2004). In particular, it fails to address the issue of working within a sociocultural system in which doctors must implement solutions and adapt standard procedures to multidimensional, dynamic, and time-critical problems (an issue also common to other domains of work practice; see Boulton & Cole, 2016).

Performance measurement. One of the challenges in the scientific study of medical decision making is the operationalization of performance quality (for a recent discussion see Feufel & Flach, 2019). Recent initiatives attempt to base physician and/or hospital reimbursement on care *quality* (value-based

reimbursement) rather than *quantity* of services (fee-for-service reimbursement) (e.g., Henkel & Maryland, 2015). Indeed, patient satisfaction is a key component of how physicians are evaluated and even reimbursed in the current U.S. healthcare system. However, patients with very similar complaints may nonetheless have very different expectations of care, and physicians must be sensitive to those differences in order to reduce complaints or prevent return visits to seek additional intervention. A great deal of applied research in medicine focuses on the detection and prevention of medical error as a measure of provider quality, spanning errors in medication dosage (Gonzales, 2010; van der Veen et al., 2018), equipment utilization (Walsh & Beatty, 2002), and, most prominently, misdiagnosis (Maude, 2014) with biased thinking and intuition receiving particular attention (Croskerry, 2013; Norman & Eva, 2010). On the other hand, the absence of erroneous outcomes is a low bar that does not assure quality care, in particular in values-driven disciplines such as medicine.

Gigerenzer and colleagues advocate a criterion of *ecological rationality*, in which decision quality is a function of compatibility with the structure of the environment (Gigerenzer, 2008; Todd & Gigerenzer, 2007). However, the environment of medicine, as in nearly all work domains, is culturally determined by values and norms (Croker et al., 2008; Foy et al., 2010; Johnson et al., 2008; Martin et al., 2010; Shalin et al., 1997) that are not necessarily coherent. Practitioners must balance competing interests, for instance, between patients and hospitals when defining treatment goals (Atkins & Ersser, 2008; Engestrom, 1993) and balance efficiency with patient safety when enacting these goals in light of resource constraints on one hand and medical standards on the other (e.g., Feufel, 2009). With the exception of some qualitative observational studies (Benner, 1982; Currey & Botti, 2003; Donner-Banzhoff et al., 2017; Engestrom, 1993; Feufel, 2009; Shalin & Bertram, 1996), the in situ decision-making *process* is rarely the measure of interest. We use multilevel modeling to quantitatively examine the effect of clinical context on decision-making processes.

Multilevel Modeling

Multilevel analysis quantitatively examines the effects of predictors on nonindependent observations involving multiple levels of analysis, such as when unique individuals (i.e., patients) are nested within and therefore dependent on a particular grouping variable (i.e., doctor). More specifically, differences between doctors would cause patients who are treated by the same doctor to experience more similar patient-related reasoning processes (e.g., Brooks et al., 1991) compared to patients who are treated by a different doctor. Multilevel modeling allows one to examine differences between doctors as well as interactions between patient-level variables and doctor-level variables. This allows us to study variance in the criterion measure (in this case, decision-making processes, measured for individual patients) separately for patient-level variables that vary across individual patients (Level 1 predictors) and variables that vary across doctors/work settings (Level 2 predictors). Multilevel analysis allows one to examine variance in the intercept (conceptually similar to main effects for variables at both the patient and doctor level) and slope (interactions) in response to a given predictor. See Supplementary Material 1 for a more detailed explanation.

We shadowed emergency physicians as they treated patients in their normal work setting and recorded observable behaviors. Splitting observations across two separate hospitals allowed us to identify specific influences of the work system. The observed individual behaviors formed the basis for aggregated criterion measures reflecting more general decision-making processes identified via factor analysis. Factor analysis distinguished between two phases of decision making, *goal establishment* and *goal enactment*. This distinction is reinforced by multilevel modeling results, revealing differences in the response of these phases to contextual influences, which served to guide qualitative analysis of our observation notes. Our interpretation was aided by a collaborating clinician who provided insight into various hospital practices, the clinical practice of medicine, and potential ways in which clinical

practice may be affected by environmental features.

Identifying the response to different hospital contexts is the main goal of the present study. We demonstrate the responsiveness of goal enactment to hospital while controlling for patient and shift difficulty, along with physician experience. The effect of hospital is therefore independent of other contextual features for this measure of reasoning/decision making. We reveal the measure goal establishment as responsive to an interaction between patient difficulty and hospital. Goal establishment behaviors increased with patient difficulty in the suburban hospital, independent of shift difficulty, whereas the relationship between patient difficulty and goal establishment was inconsistent in the urban hospital. Qualitative results illustrate these quantitative patterns.

METHOD

Participants and Recruitment

This study was approved by IRBs at the teaching hospitals under observation as well as the university with which the medical school was affiliated (described below). Nine second-year residents, eight third-year residents, and nine attending physicians were shadowed for this study. The study was introduced to the physicians at each hospital via emails and announcements at lectures and staff meetings, facilitated by a clinician collaborator on faculty at the medical school. Physicians were eligible for inclusion if they worked in the emergency department of one of the hospitals under study and were beyond the first year of residency. As we were concerned that patient population and complaints were likely to differ over time (i.e., the distribution of complaints seen late Friday night is likely to differ from the distribution of complaints seen on a Tuesday morning), we attempted to control for temporal confounds by counterbalancing observations across hospital, part of the week, and time of day. Physicians were selected for recruitment based on compatibility between the monthly shift schedule and our counterbalancing scheme. Identified physicians were emailed to further explain the study and assess interest in participating.

Observation Setting

Observations occurred over a period of 18 months. We conducted a total of 26 observation sessions across two different teaching hospitals associated with the medical school of a single public Midwestern university. One hospital was a suburban ED with approximately 40 beds, serving patients that were primarily older, Caucasian, and insured. The suburban hospital used a paper-based record system including "t-sheets," which served to document the patient interview and care process. The other hospital was an urban ED with approximately 60 beds that served patients that varied in age, with a larger percentage of minority and uninsured patients. The urban hospital used an electronic record system.

Procedure

Participating physicians were shadowed by a single observer for the duration of one work shift (generally lasting 10 hr). The observer took handwritten notes of all directly observable actions or thoughts articulated by the physician, using a stopwatch for timestamps. Sensitive exams such as rectal or pelvic exams were not observed. Information that could be used to identify patients or doctors was not recorded at any time. When convenient, the observer asked general questions such as "what are you thinking about this patient?" or "do any lab values jump at you?" to avoid potentially leading the doctor. In the handful of cases where a doctor asked the observer to get a blanket or something similar, this was noted in the observation notes and coded as if the doctor had asked a staff member to do the same or had done it themselves. Patients gave verbal consent to have the observer present during interviews after receiving an explanation of the study from the doctor who was being observed. No patients refused to allow the observer to be in the room for interviews. Patients who did not receive an explanation or were unable to consent were excluded from the analysis, yielding 239 total patient encounters.

Analysis

Our predictors and criterion measures were each derived via analysis of the observational

data. Predictor variables reflected situational/contextual factors (e.g., the hospital where the observation occurred), whereas criterion measures were computed based on counts of specific doctor behaviors coded in the observation notes. The predictor variables are described first, followed by a description of the variables that contributed to our criterion measures. Computation of the criterion measures was based on behavioral processes identified via exploratory factor analysis; criterion measure scoring is therefore described in the Results section.

Predictor variables. To investigate context effects on medical decision making, we included differences between patients (e.g., patient difficulty), hospital settings (e.g., hospital-specific work practices), and doctors as predictor variables. We elaborate each type of predictor variable in turn.

Patient-level predictors. We used a measure of *patient difficulty* to capture the influence of individual patients on physician behaviors and facilitate operationalization of overall workload during a shift (described below). In contrast to familiar triage status measures (e.g., the Emergency Severity Index; Gilboy et al., 2011), we conceptualized patient difficulty not strictly as a measure of medical complexity or urgency, but as a more general measure of how much a patient contributed to the total workload of a shift, relative to other patients. We used a 3-level measure of patient difficulty (1 = minor effort, 2 = average effort, 3 = major effort) that captured more than the actions associated with collecting histories, performing physical examinations, or administering treatments by including elements of care such as communication and coordination and patient cooperativeness.

We attributed a difficulty score for each patient based on a review of the observation notes to generate a subjective assessment of the amount of effort and time invested by the doctor for each patient seen during a shift, incorporating traditional factors such as workups and interventions, but also accounting for time spent treating/dispositioning the patient, whether consultations with other physicians were required, how typical or atypical the physician judged a patient to be, how cooperative the patient was, and so on. Patients who required little effort and minimal workup and/or presented with very straightforward medical problems (interestingly, labeled in ED work practice as "treat and street") received a 1. Examples of complaints that generally (though not necessarily) received a 1 include sinus infections and cuts requiring sutures but no other intervention. Patients who required some effort, more attention, and/or greater workup/intervention with limited complexity (the average patient; e.g., a "textbook" presentation of shortness of breath or chest pain after exertion in a patient with a history of heart problems) received a 2. Difficult/uncooperative patients or patients with complicated complaints that required a large amount of the doctor's time or attention received a 3. Patient difficulty was tested for reliability by recoding all of the patients from six of the observed shifts. Weighted Cohen's Kappa was 0.73, indicating reliable ratings. Based on the rating definitions, the "average" patient was supposed to be scored as 2. The mean patient difficulty score across all analyzed patients was in fact 1.98.

Doctor-level predictors. Three predictors differed between doctors: the difficulty of the doctor's shift, the doctor's years of clinical experience, and the hospital in which the doctor worked.

Shift difficulty. Workload affects clinical reasoning processes, including how practitioners prioritize tasks, select interventions, and interact with patients (Smith et al., 2008). As patient care is the physicians' primary responsibility during a shift, aggregated individual patient difficulty scores served as the basis for calculating a total shift difficulty score. In the case that a patient was outside of the direct care of the doctor being shadowed (i.e., an attending physician interacting with patients seen primarily by a resident under their supervision), that patient's difficulty score was not included in the shift difficulty score. The shift difficulty score was therefore computed as the sum of the difficulty scores of the patients under the doctor's direct care during the shift. Shift difficulty scores ranged from 13 to 44.

Experience. Experience and associated expertise affect how people process information, make decisions, and respond to

environmental stressors (Arora et al., 2010; Dreyfus & Dreyfus, 2005). We operationalized experience based on the number of years a physician had been in clinical practice. Second- and third-year residents were scored to have 2 and 3 years of experience, respectively. Each attending physician's experience was calculated based on the year in which the doctor graduated from medical school. The attending physicians had between 5 and 31 years of experience, with a mean of 13.2 years.

Hospital. The social and physical setting in which one works can have a dramatic influence on behavioral processes (Hutchins, 2014). We therefore included the hospital where the doctor worked, each with its own unique work system, as a nominal variable.

Criterion measures. The criterion measures for this study corresponded to the observable behavioral processes of emergency physicians. Specific behaviors of interest were identified a priori based on our research questions and literature review, as well as grounded in actual practice via participant observation in weekly lectures for the residents over a period of approximately 6 months. The coding scheme was refined further via testing on pilot observation notes.

The final coding scheme distinguished between various aspects of behavior related to patient care and disposition in the emergency department (Table 1; see Supplementary Material 2 for the full coding manual). *Information gathering* behaviors were used by the doctors to generate facts about the case at hand, coded based on both the source and type of information. *Diagnostic* behaviors were intended by the doctors to establish a cause for symptoms or eliminate potential causes from consideration. *Evidence evaluation* behaviors pertained to the evaluation of the completeness of one's understanding of the situation or the quality of the available evidence. *Patient management* behaviors were used to treat or otherwise manage patients during their stay in the emergency room. *System management* behaviors aided the doctor in working within the constraints of the hospital or the larger healthcare system. *Filtering* behaviors involved limiting the problems that the doctor had to address and determining the scope of the patient's complaint.

TABLE 1: Listing of the Coded Behavioral Variables.

Category	Subcategory
Information gathering	
Type	
	Current symptoms
	Timeline
	Past medical history
	Contributors
	Reference information
	Other
Source	
	Exams
	Patient
	Tests/images
	Patient's family/friends
	Medical records
	Hospital staff
	Internet/reference material
	Miscellaneous
Diagnostic behavior	
Evidence evaluation	
Patient management	
	Treatment
	Consulting
	Collaboration
	Logistics
System management	
Filtering behavior	

Once coding was completed, an iterative, exploratory factor analysis served to reduce the numerous coded aspects of physician behavior across the sample of patients into interpretable groups of behaviors representing higher-order constructs. This empirically driven approach to construct identification generalizes over specific patient complaints to reveal more general patterns of physician behavior in an uncontrolled setting (see Donner-Banzhoff et al., 2017, for

a similar approach). Factor scores, which were calculated based on the frequency of behaviors within each factor, served as criterion measures in our multilevel analyses.

Reliability check. The first author served as the coder for all of the observational notes. Five months later, the same coder re-coded six of the observed shifts to assess code–recode reliability (the same subset of shifts was used to check the reliability of the patient difficulty score). We compared re-coded observations to the original codes and computed the number of instances of agreement for both observations, instances of affirmative codes for the original but a failure to code a behavior the second time, instances in which a behavior was not coded the first time but coded the second time, and instances in which a behavior was not coded in both observations. These categories were summed across all six of the observed shifts in order to ensure adequate sample sizes. Reliability for each variable was assessed individually using Cohen's Kappa. The totals for each category were entered into a 2 × 2 contingency table for analysis. Cohen's Kappa values for the variables indicated that coding was sufficiently reliable (Table 2). We note that some Kappa values can likely be attributed mainly to the high proportion of correct rejections. Nevertheless, we felt it was important to account for correct rejection in addition to correct identification.

TABLE 2: Cohen's Kappa Values for the Variables in the Analysis

Variable	Hit/Hit	Hit/Miss	Miss/Hit	CR	Kappa	*SE*	95% CI Lower Bound	95% CI Upper Bound
Patient	1479	78	97	2618	0.91	0.01	0.90	0.93
Contributors	235	27	55	3501	0.84	0.02	0.81	0.87
Evidence evaluation	404	125	103	3319	0.75	0.02	0.72	0.78
Diagnostics	279	41	52	3486	0.84	0.02	0.81	0.88
Tests/images	364	87	43	3484	0.83	0.02	0.80	0.86
System management	410	86	47	3258	0.84	0.01	0.81	0.87
Collaboration	1021	112	70	3092	0.89	0.01	0.87	0.91
Logistics	168	134	25	3458	0.66	0.03	0.61	0.71
Internet/reference	14	1	1	3750	0.93	0.05	0.84	1.00
Reference information	14	1	1	3750	0.93	0.05	0.84	1.00
Exam	413	16	11	3459	0.96	0.01	0.95	0.98
Family/friends	99	8	21	3675	0.87	0.02	0.82	0.92
Medical records	250	33	49	3575	0.85	0.02	0.82	0.88
Staff	333	27	56	3586	0.88	0.01	0.85	0.90
Miscellaneous info source	28	26	31	3690	0.49	0.06	0.37	0.60
Current symptoms	1705	196	141	2496	0.85	0.01	0.83	0.86
Timeline	63	16	9	3683	0.83	0.03	0.77	0.90
Past medical information	684	50	112	3163	0.87	0.01	0.85	0.89
Other information	185	90	81	3457	0.66	0.02	0.61	0.71
Consult	73	51	8	3637	0.71	0.04	0.63	0.78
Treatment	250	36	46	3537	0.85	0.02	0.82	0.88
Filtering	25	27	13	3704	0.55	0.06	0.43	0.68

Note. CR = correct rejection (variable was not coded in either instance).

Some behaviors were coded more reliably than others. This is particularly true for behaviors such as filtering and using miscellaneous sources of information. The variables that were least reliable eventually dropped out of the analysis (see the results of the factor analysis below).

RESULTS

Our analyses followed several steps combining both quantitative and qualitative approaches (e.g., Cummings, 1980). The following results begin with a summary of an exploratory factor analysis used to aggregate the large number of individually coded doctor behaviors into higher-level factors that potentially distinguish doctors from one another: goal establishment and goal enactment. We next describe the aggregation of behaviors within each factor into a single factor score, which served as a criterion measure for subsequent analyses. We then describe the results of multilevel modeling used to identify variation in these two factor scores related to patient difficulty, shift difficulty, hospital, and physician experience. The effect of experience on behavior is a substantial research area in its own right; here we omit results related to differences in physician experience and focus on the general contextual influences on physician decision making.

The results of the multilevel analysis served as a guide for identifying and interpreting excerpts in the observational data. Illustrative excerpts from the observational notes are included with the quantitative results for each factor (e.g., Ho et al., 2017). Hypothesized causes of our results and implications of our findings are addressed in the discussion section.

Factor Analysis for Aggregating Behavior Into Criterion Measures

An iterative, exploratory factor analysis was used to reduce the numerous coded aspects of physician behavior into interpretable groups of behaviors that could be used as a criterion measure for multilevel analysis. Six outlier patients (exceeding ±5 standard deviations from the mean on any single behavior) were excluded from the analysis, resulting in a final sample of 233 patients. The initial factor analysis yielded a total of seven factors with eigenvalues greater than the standard cutoff of one. A scree plot supported a three-factor model. We only retained a contributing doctor behavior if its loading exceeded 0.5 on a single factor and the difference in that behavior's loading across multiple factors exceeded 0.35 (with flexibility of a couple of hundredths to allow for conceptually convincing variable inclusion). We conducted multiple iterations of the analysis, using both criteria to remove variables until all remaining behaviors met the loading criteria. Compared to models retaining two or one factor(s), the three-factor model ultimately emerged as the most conceptually sound and is presented here.

Retained three-factor model description. Factor 1 in the final version contained action-oriented management behaviors (i.e., *evidence evaluation behaviors, diagnostic behaviors, system management behaviors, using tests and images to gather information, logistic behavior*, and *collaborative behaviors*), hereafter referred to as the "goal enactment factor." Factor 2 contained behaviors that indicated a recognition of uncertainty (i.e., *using texts or the Internet* to find *reference information),* hereafter referred to as the "uncertainty reduction factor." Factor 3 consisted of gathering information from various potential sources, primarily related to the patient interview (i.e., finding out about *contributing factors* and *using the patient as a source of information*). Because such behavior occurs primarily at the outset of the session (Roter & Hall, 2006), we refer to this as the "goal establishment" factor. In subsequent analysis, we were unable to explain variance in the uncertainty reduction factor; it is therefore not discussed further to save space.

Deriving criterion measures. Factor analysis supported the aggregation of low-level variables into factor scores for the two retained factors as follows. For every patient, each variable within a factor was converted into a Z score and then averaged together with the Z scores of the other variables in a factor to obtain an aggregated factor score for each patient. These factor scores served as criterion measures in

multilevel analyses for examining differences in patient treatment, including variables related to the patient, the doctor, and the environment.

We broadly structure our exposition of the multilevel analysis by the two aggregated behavioral measures revealed in factor analysis: goal establishment and goal enactment, and the effects of predictors within these measures. The multilevel analyses followed the model-building steps recommended by Bliese (2002). An alpha level of 0.05 (truncated) was adopted for all analyses. In general, goal establishment and goal enactment respond to hospital context differently, simultaneously distinguishing between these constructs and clarifying the nuanced influence of context on decision-making behavior. We follow up with example cases drawn from the observations to illustrate these findings.

Contextual Effects on Goal Establishment

Context affects goal establishment as *an interaction between patient difficulty and hospital*. Simply, physicians responded to increases in patient difficulty differently between hospitals. We established this finding as follows.

The intraclass correlation coefficient (ICC) for the goal establishment factor was 0.41, indicating substantial between-doctor variance on this factor. Patient difficulty positively predicted goal establishment scores. We next attempted to use doctor-level variables (i.e., experience, shift difficulty, and hospital) to account for variance in the intercept. Experience was negatively related to goal establishment behaviors. Shift difficulty and hospital did not predict these behaviors (Table 3). A deviance chi-squared test indicated that there was little variability in the relationship between patient difficulty and goal establishment across physician-level predictors, $\chi^2_{diff}(2) = 0.87$, $p > 0.05$. Because this test for slope variance has low power (LaHuis & Ferguson, 2009), we continued by examining whether patient difficulty was more strongly related to decision-making behavior for some doctors than for others. As seen in the "slope" section of Table 3, hospital significantly predicted the relationship between patient difficulty and goal establishment behaviors such that their relationship was stronger for doctors in the suburban hospital than in the urban hospital (Figure 1). Doctors in the suburban hospital increased their goal establishment behaviors in response to increasing patient difficulty whereas doctors in the urban hospital demonstrated inconsistent responses to increasing patient difficulty. The other doctor-level predictors (experience and shift difficulty) did not affect the statistical relationship between patient difficulty and goal establishment behaviors. However,

TABLE 3: Multilevel Results for the Goal Establishment Factor

Model and Parameter	Parameter Estimate	*SE*	*T*	*df*	*p*
Level 1					
Patient difficulty*	0.16	0.07	2.21	228	0.03
Level 2					
Intercept					
Experience*	−0.04	0.02	−2.34	22	0.03
Shift difficulty	−0.01	0.02	−0.78	22	0.44
Hospital	0.04	0.22	0.17	22	0.86
Slope					
Experience	0.02	0.01	1.54	22	0.14
Shift difficulty	−0.02	0.01	−1.94	22	0.07
Hospital*	0.43	0.16	2.66	22	0.02

Note. Predictors with an * are significant at $p < 0.05$.

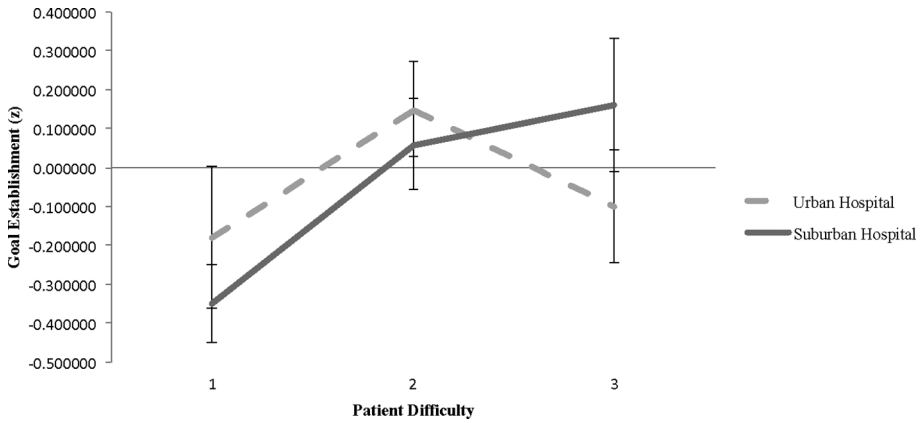


Figure 1. Graph of the slope effect of hospital on the goal establishment factor. *Note*. Error bars represent standard error.

because of its potential policy implications and *p* values that did not conclusively rule them out, we provide qualitative illustrations of the possible influence of shift difficulty.

Hospital*patient difficulty effect on goal establishment. The following examples illustrate the differences in goal establishment behaviors with increasing patient difficulty in each hospital (Figure 2). For comparison, the example from the suburban hospital is from a patient treated by an attending physician with 13 years of experience during a shift with a difficulty score of 30. The example from the urban hospital is from a patient treated by an attending physician with 15 years of experience during a shift with a difficulty score of 19. Both patients presented with similar complaints of chest pain.

Goal establishment behavior is primarily found in the patient interview; differences across interviews speak to potential influences on goal establishment. The doctor in the suburban hospital asked her patient more details about the patient's past history and other information (lines 23–51) and focused less attention on the patient's current experience of her symptoms (lines 1–8). In contrast, the doctor in the urban hospital asked fewer questions about the patient's past history but spent more time asking about the patient's current symptoms (lines 9–20). The doctor in the urban hospital also acquired a greater proportion of information from sources *other* than the patient such as the medical record. The doctor at the suburban hospital mainly got information from the patient and rarely used the medical record. Both physicians gathered similar information, but importantly went about it in different ways. We suggest an explanation for this pattern in the discussion.

Potential shift difficulty effect on goal establishment. The statistical analysis does not rule out an effect of shift difficulty ($p = 0.07$). Because this has important policy implications, we examined the related qualitative data. Emergency physicians are under constant pressure to do things quickly, and they must obtain information from their patients efficiently. Both residents and attending physicians voiced the effect of time on their practice:

> *Resident:* Internal medicine and emergency medicine are very different. In internal medicine you can go back hours later and ask the patient something, and you can take a 25 min history. In emergency medicine you have 5 min to get what you can, and you can never get everything (you just can't)...
>
> *Attending:* I like it when it's slow [not many patients to see] because I can talk to the patient and get their story rather than

		Suburban Difficult Patient		Urban Difficult Patient
Line	Speaker	Utterance/action	Speaker	Utterance/action
1	*Attending*	What's wrong?		*The attending speaks to the EMS squad that brought the patient in. They tell him the patient has chest pain, is sinus tach, is hypertensive, O2 sat is 100%, the patient has no edema and has had recent lasix and pneumonia.*
2	*Patient*	I have chest pain.	*Attending*	So you have congestive heart failure, high blood pressure, and your heart is racing?
3	*Attending*	Does anything bring it on?	*Patient*	Yeah.
4	*Patient*	No, it just does its thing.	*Attending*	Did you take lasix today?
5	*Attending*	What does that mean?	*Patient*	No, I was never told about lasix.
6	*Patient*	It hurts and I have to lie down.	*Attending*	Who is your primary doctor?
7	*Attending*	Does that help it?	*Patient*	I don't have one.
8	*Patient*	No I just rest and wait for it to subside. It hurts in different ways – like a fluttery, prickly feeling.		*The attending listens to the patient's back.*
9	*Attending*	Have you seen your doctor this week?	*Attending*	Do you have any chest pain?
10	*Patient*	No.	*Patient*	I had back pain earlier.
11	*Attending*	Why not?	*Attending*	Is that chronic?
12	*Patient*	It doesn't seem to help.	*Patient*	I had an accident years ago.
13	*Attending*	Has this happened before?	*Attending*	Are you short of breath?
14	*Patient*	Not to this extent.	*Patient*	Yes.
15	*Attending*	What changed today that brought you in?	*Attending*	What about belly pain?
16	*Patient*	I got a crying sensation, like I just wanted to cry.	*Patient*	No.
17	*Attending*	Do you have that right now, or has it gone away?	*Attending*	Did you pass out?
18	*Patient*	It's still there. I can't describe it.	*Patient*	No.
19		*The attending listens to the patient's back.*	*Attending*	Are you dizzy?
20	*Attending*	Who's your doctor?	*Patient*	Yes. I get short of breath when I go to the bathroom.
21		*The patient tells the attending her doctor's name; her husband adds who the patient's cardiologist is.*	*Attending*	So this happens every day?
22		*The attending listens to the patient's chest.*	*Patient*	Yes. I was in the hospital for two weeks last month.
23	*Attending*	When did you see your doctor last?		*The attending listens to the patient's chest.*
24	*Patient*	A couple of weeks ago.	*Attending*	Your legs aren't very swollen.
25	*Attending*	Did you get any medications?	*Attending*	We'll get an xray and blood work but I don't hear any crackles.
26	*Patient*	I didn't take them after the first day because I didn't notice a change and I'm on a lot of meds already.	*Attending*	Do you smoke?
27	*Attending*	Why were you given the meds – the same thing?	*Patient*	I quit when I went to the hospital last month. I have a dry cough, too.
28	*Patient*	Yeah.	*Attending*	Do you have a breathing machine at home?
29		*The attending checks the patient's pulse in her foot.*	*Patient*	No.
30	*Attending*	Do you have any asthma, diabetes, cancer, or stoke history?	*Attending*	It may be COPD and not heart failure but we'll check your heart and look at some stuff. We'll probably admit you.
31	*Patient*	No, nothing.		
32	*Attending*	Has anything been off in the last few days like feeling dizzy or headaches or diarrhea?		
33	*Patient*	No. The pain goes to my back sometimes. I also have a stent. I went to the doctor because I got the crying sensation again then I got the stent.		
34	*Attending*	Have you had a recent stress test?		
35	*Patient*	I had a chemical stress test two weeks ago.		
36	*Attending*	Did they find anything?		
37	*Patient*	No.		
38	*Attending*	Have you had any bowel, bladder, or weight changes?		
39	*Patient*	No.		
40	*Attending*	What about family history?		
41	*Patient*	Nothing except for leukemia in my sister.		
42	*Attending*	How did your parents die?		
43	*Patient*	They were just old.		
44	*Attending*	Have you started menopause or had a hysterectomy?		
45	*Patient*	Hysterectomy.		
46	*Attending*	Have you had any other surgeries?		
47	*Patient*	I had my tonsils out.		
48	*Attending*	Are you taking any herbal supplements or anything like that?		
49	*Patient*	No.		
50	*Attending*	Have you had anything with your appendix or gallbladder?		
51	*Patient*	I still have them both.		
52	*Attending*	I'll check your med list and get some tests and we'll see what's going on. Do you have any other doctors?		
54	*Patient*	No.		
55		*The patient's husband gives the attending her medication list.*		
56	*Attending*	I'll try to get the meds from the doctor too and find out about them.		

Figure 2. Example of similar patients treated at different hospitals.

TABLE 4: Multilevel Results for the Goal Enactment Factor

Model and Parameter	Parameter Estimate	*SE*	*T*	*df*	*p*
Level 1					
Patient difficulty*	0.59	0.05	11.11	228	<0.01
Level 2					
Intercept					
Experience	−0.02	0.01	−1.46	22	0.16
Shift difficulty	−0.01	0.01	−0.64	22	0.53
Hospital*	−0.40	0.15	−2.59	22	0.02
Slope					
Experience	0.01	0.01	0.52	22	0.61
Shift difficulty	−0.01	0.01	−1.43	22	0.17
Hospital	−0.14	0.15	−0.97	22	0.34

Note. Predictors with an * are significant at $p < 0.05$.

> just try to get what I need. Everyone has a story and I like to hear them.

In the statements above, the resident indicates the sense of time pressure and implies that interviews become shorter due to being busy, while the attending physician limits questions to only the most relevant information during busy shifts. Both describe ways in which they are affected by the need to accommodate high patient volume.

Contextual Effects on Goal Enactment

Context affects goal enactment as main effects of patient difficulty and hospital (i.e., physicians performed more enactment with more difficult patients and, independently, more enactment in the urban hospital). We established this finding as follows.

The ICC for the goal enactment factor was 0.32, again indicating substantial between-doctor variance on this factor. As with goal establishment, physicians demonstrated more goal enactment behaviors with more difficult patients (Table 4). We next attempted to account for variance in the intercept using doctor-level predictors. The hospital where the doctor worked predicted variance in the intercept such that doctors at the urban hospital performed more goal enactment behaviors. This effect is different than the hospital effect for goal establishment, in which hospital accounted for slope variance. Experience and shift difficulty did not predict intercept variance in the goal enactment behaviors. A deviance chi-squared test demonstrated that allowing the slopes to vary improved the model fit, $\chi^2_{\text{diff}}(2) = 10.16$, $p < 0.01$. Our analysis of slope variance revealed that in contrast to goal establishment, hospital and shift difficulty did not interact with patient difficulty and goal enactment behaviors.

We now consider the qualitative support for the quantitative effects of patient difficulty, followed by the effects of hospital. As with goal establishment, we complete the review of goal enactment effects by exploring possible effects of shift difficulty.

Effect of patient difficulty on goal enactment. Patient difficulty predicted goal enactment behavior such that doctors performed more goal enactment behavior with more difficult patients. Intuitively, more complex cases should require more tests and care. Recall, however, that patient difficulty was not based solely on medical complexity. The following example from a patient with a headache and nausea/vomiting provides some insight:

> *Attending (to the observer upon leaving the room):* I don't think there's anything really wrong with the patient. Just gastroenteritis, but the patient will be hard to deal with because she's a tough stick [hard to find a vein for injections or blood draws]. She'll be hard to discharge. I'll get labs and an x-ray because the patient will elevate her symptoms [complain about new things to get what she wants] if I don't. An upfront and thorough workup will be more acceptable to her (that's my gut feeling), so she'll be ok and leave.

This patient was not very difficult from a medical perspective, but the attending physician believed that the patient would resist being discharged and complain without a thorough workup. As a result, the attending physician invested more energy in discharging the patient than an objective medical assessment would require.

Effect of hospital on goal enactment. In goal establishment reviewed above, hospital affected the *relationship* between patient difficulty and goal establishment. In contrast, doctors at the urban hospital simply *did more enactment* than doctors at the suburban hospital (see intercept analysis of Table 4). The excerpts below from observations at the urban hospital illustrate ways in which the patient population and work practices may affect decision-making processes:

> *Attending:* There's a difference in decision making between here [the urban hospital] and [the suburban hospital]. Follow up over there is more reliable, and there are different patient expectations.

The following example illustrates the types of actions doctors at the urban hospital sometimes had to perform that doctors at the suburban hospital typically did not. In this case, an intoxicated patient reported to the ED with a cut after falling:

> *Observer notes:* The attending sees that the head CT is ok, but the face CT shows fractures (the attending had not been expecting that).
>
> *Attending:* The eye muscle may be trapped—the patient will need plastics and ophthalmology. I'll have to call for the patient. She's homeless so she probably won't call on her own. I'll need to hold her until she's sober and explain the injury. I'll call plastics in the morning.
>
> *Observer notes:* The attending arranges for the charge nurse to hold a room somewhere for the patient until she's sober. The attending explains the situation and apologizes for keeping a room for so long. The attending calls the clinic and speaks with a plastics resident. The attending explains the patient's case to be sure that follow up is ok. The attending says that the patient is homeless so they'll keep her in the ED and send her over to the clinic at noon. The attending tells the observer that she's trying to get good follow up. She says that it will save a re-visit later and it will make sure that the patient gets follow up care (the patient is uninsured and doesn't really have any family).

This somewhat extreme case demonstrates the types of issues that can arise with patients at the urban hospital that are less likely to arise at the suburban hospital. This was not an imminent risk such as a stroke or an aneurism that would appear as demanding from a triage-based assessment, but the case required substantial enactment effort. The patient lacked the resources to arrange her own follow-up care, so the attending physician had to use valuable hospital resources (a bed in a busy ED), took time to personally arrange follow-up care at a local clinic, and took steps to be sure the patient would be able to get there (the clinic was adjacent to the hospital). A medically similar patient at the suburban hospital may not require the same effort.

In addition to patient population, work practices differed across hospital with implications for goal enactment. One practice that differed across hospitals was that triage nurses in the

urban ED would commonly order lab tests or x-rays before the patient was actually seen by the doctor in order to minimize the amount of "down" time spent waiting for results to come back. The following example demonstrates how this practice may increase enactment behavior:

> *Observer notes:* The resident sees that the triage nurse had ordered x-rays and cardiac labs—the resident wasn't going to order them. An attending physician comes by and says that some labs are great to order in triage most of the time, but some labs should only be ordered by a doctor so the doctor can determine if they're really needed because *labs can send the doctors down a path that they're obligated to investigate*. The resident agrees.

By ordering tests independently of the physicians, the triage nurses often saved time, but sometimes risked creating additional work by forcing the doctors to follow up on abnormal but irrelevant values or document why they chose not to do so. Such work practices affect goal enactment behaviors independently of anything related to the patient's complaint.

Potential shift difficulty effect on goal enactment. As with goal establishment, we cannot rule out an effect of shift difficulty ($p = 0.17$). This exchange between an attending physician and an admitted patient's family offers an example of physicians specifically altering their diagnostic behavior (a component of goal enactment) in order to adapt to a busy shift:

> *Attending*: Do you have any questions?
>
> *Patient's family*: Yes; The nursing home mentioned there were drugs in the urine?
>
> *Attending*: We don't care about things like that [illicit drugs] in the ED.
>
> *Patient's family*: We care because of the [organ] donor list.
>
> *Attending*: Oh, ok.
>
> *Patient's family*: Could it be from meds at the nursing home?
>
> *Attending*: Maybe, could be Vicodin or something else.
>
> *Patient's family*: Can you find out?
>
> *Attending*: Ask about it upstairs [the patient is being admitted], we're too busy here to find out now.

In this case, a drug screen had apparently already been performed and presumably would have been available to the physician. However, the physician in this case had already admitted the patient to the hospital for additional care, so he decided not to follow up on the prior screening at least partly because he was busy (and partly because the presence of drugs is not necessarily an issue requiring attention as far as emergency medicine is concerned). The attending physician did take the time to ask if the family had any questions and explained what the drug may have been, however.

DISCUSSION

Our interest in contextual influences on decision-making processes necessitated a rather substantial departure from standard experimental methods and measures, applying multilevel modeling to observational data of physicians in their natural setting. Multilevel modeling revealed two key findings specific to the effect of contextual influences on decision-making processes. First, we demonstrated the responsiveness of goal enactment to hospital independent of patient and shift difficulty. Patients received *more enactment care in the urban* hospital. Second, the observed ED physicians appeared to tailor their goal establishment response to patient difficulty broadly construed as an interaction with hospital, such that more *difficult patients received more establishment care in the suburban hospital*, which we also demonstrated independent of shift difficulty. In so doing, we illustrated the importance of accounting for contextual considerations in the study of medical reasoning and identified issues

to consider in the effort to measure care quality. We first consider these findings with respect to possible causal factors. We complete our discussion with the theoretical, methodological, and practical implications of our work, along with limitations and future research.

Contextual Influences on Goal Establishment and Goal Enactment

Factor analysis on observable behavior revealed two separable but interrelated decision-making processes in the emergency department. Physicians use *goal establishment* behaviors to identify patient needs and desired outcomes, and *goal enactment* behaviors to implement solutions within the sociocultural constraints of emergency medical practice. The effect of hospital and patient difficulty differs on these measures and supports their distinction as constructs. We discuss the patient difficulty effects first and then turn to the effects of hospital and shift difficulty.

Patient difficulty. More difficult patients elicited more goal enactment behaviors. However, in our study, medical complexity was not the sole criterion for assigning a patient difficulty score. Quality care requires physicians to identify patient demands and balance those desires against genuine medical need in order to justify their actions and limit cost while maintaining patient satisfaction. Emergency physicians must coordinate with other physicians to admit or discharge patients, order and review tests (sometimes iteratively), and meet with patients and their families throughout the care process. These additional duties mean that a patient's demands on a doctor's time are not solely dependent on the patient's medical complaint. These doctors balance multiple, sometimes conflicting, goals such as economic and medical standards along with the patient's needs and preferences, all while respecting the constraints imposed by the resources available to resolve particular problems.

More difficult patients also increase goal establishment behavior, albeit in a more nuanced fashion. Unlike enactment, we showed that this general effect on establishment is dependent upon the hospital, discussed separately below.

Hospital. The hospital in which the doctors worked had differing effects on goal enactment and goal establishment, as a main effect or interaction with patient difficulty, respectively. The hospital setting affects enactment behaviors such as ordering tests and images or system management behavior by altering the doctors' work structure and work constraints (e.g., the availability of electronic records or the frequency of labs being ordered in triage). The generally higher number of patients may have led the urban ED to adopt work practices such as ordering more labs in triage in order to maximize their patient throughput. That is, the number of tests that are ordered says at least as much about the hospital as it does about the patient, and illustrates the potential risk of relying on decontextualized measures (e.g., tallies of interventions) to assess workload or patient need. The tradeoffs involved with ordering a greater number of labs in triage also illustrates the potential for competing values (the need to see patients quickly vs. the need to minimize cost) to influence not only physician behaviors, but the practices of the work system as a whole.

Differences in patient populations between hospitals can also affect goal enactment behaviors. The higher proportion of uninsured and low socioeconomic status patients at the urban ED sometimes forced the doctors to engage in logistic behaviors such as finding free clinics or cheaper medications for their patients. Hospital effects on goal establishment are more nuanced. The suburban hospital showed a consistent increase in goal establishment with patient difficulty, but the urban hospital did not. Differences in patient population may account for this interaction. Suburban patients may have different perceived expectations of care.

An additional possible explanation for the interaction is related to the role of technology in the work system of each hospital. The urban hospital used an electronic records system, whereas the suburban hospital used a primarily paper-based system (some records were electronic). The suburban hospital's records system lacked easy access to a patient's history, forcing

the doctors to rely more on the patient to relay their histories and promoting more direct patient engagement.

Shift difficulty. Workload varied locally across individual patients, but also globally across shifts as a whole. Doctors are generally expected to treat each patient individually and offer the required aid without regard for what has happened with another patient. Our description of physician response to workload at both the patient and shift level indicates that this expectation is not valid. Other researchers have also examined the effect of surrounding workload on diagnostic processes (Brooks et al., 1991) or patient management (Smith et al., 2008). We expand the scope of prior researchers by describing how adjustments to workload vary across different work settings (i.e., hospital).

Theoretical Contributions

Our main theoretical contributions are in the spirit of expanding how effective medical decision making is conceptualized and measured. Doctors form and prioritize broader goals beyond diagnosis in order to guide future action and must be able to act in the world to achieve their aims. We note that the factor analysis grouped diagnostic behavior with goal enactment rather than goal establishment, indicating that the study of diagnostic reasoning does not completely capture goal setting. By highlighting goal establishment as well as goal enactment, we emphasize previously understudied aspects of patient care consistent with the analysis of expertise in other domains (Ward et al., 2016). Context-based differences in decision-making processes during nominally similar cases soften a distinction between "routine" and "nonroutine" problems, complementing Weiss and Shanteau's (2014) claim that the treatment of similar problems is the defining characteristic of expertise. Rather than focus exclusively on the novel or rare event, we believe it is equally important to study decision making across varied instantiations of similar problems.

We suggest that a critical feature of decision-making performance is the ability to adapt effort to local constraints in order to best manage available resources. Physicians in our study adapted their decision-making processes to the work setting to account for local variation in patients and potentially workload, as well as global variation in work practice. The implication of our findings is that doctors selectively and adaptively modify their behavior within disciplinary bounds in order to cope with temporal or other demands. By identifying relevant contextual features, we are better positioned to capture influences on decision making that generalize across domains and settings. Consistent with Ward et al. (2016), we emphasize that the ways in which experts manage available resources is a relatively understudied aspect of reasoning, but appears critical to the success of the physicians studied here.

Methodological Contributions

Our mixed methods including multilevel modeling employed measures of decision-making behavior that generalize across cases and enabled a quantitative treatment of observational data despite varying case content. We also reconceptualized patient difficulty. The premise of this study was that nonmedical contextual features affect physician behavior. We therefore did not expect measures such as the patient's diagnosis or triage status considered in isolation to be a reliable indicator of how much overall effort the patient required of the doctor. For example, a patient with obvious signs of a stroke may simultaneously have an urgent medical complaint but a very straightforward course of care in the ED (i.e., assess the patient, acquire the necessary labs and scans to support the diagnosis, and then admit the patient). We deliberately avoided operationalizing patient difficulty based on the complexity of a patient's complaint or measures of patient urgency such as triage scores. We instead utilized a more holistic measure of patient difficulty that reflected the totality of the demands placed on the doctor. Capturing the social aspects of patient interaction as well as the work required to manage patients within a broader work system provides a more complete sense of doctors' workload and emphasizes the importance of nontechnical aspects of patient care.

Practical Implications

Controlling the cost of medical care is an important practical concern, and variability in cost between hospitals is a key red flag (Kliff & Keating, 2013). While location-specific cost-of-living issues and hospital characteristics such as privatization are surely relevant, our findings point to more subtle albeit potentially correlated influences. Doctors may treat patients in different ways due to the systematic constraints of the hospital work system, or even the temporary constraints of individual patient characteristics and potentially the flow of the overall shift. We do not interpret our findings as an indication of flawed, biased, or intuitive reasoning to be corrected, nor do we interpret all variability as wrong.

The various ways that hospital affected reasoning processes have important implications for recent efforts in the United States to base reimbursement rates on quality of care rather than volume of services. Common diagnostic codes mask the fact that physicians are simply not performing the same task in different hospitals. Differences in physician behavior based on patient population or work systems challenge the attempt to determine hospital evaluation criteria when comparing costs or readmission rates. Technology, exemplified here by electronic medical records, does not necessarily reduce overall workload, as patients at the urban hospital still required more care. Although variability in care will never be eliminated, by understanding contextual effects on physician workload and behavior, researchers and medical educators can begin to manage care variability in predictable ways.

Limitations

Observational studies are typically accompanied by certain limitations. First, consistent with IRB considerations, notes were taken by hand without objective recording devices. Consequently, events sometimes happened faster than the observer could record them and data would be missed (e.g., questions asked during a patient interview). We have no reason to believe that missed data varied systematically between doctors or patients. Similarly, the resolution of observations was limited. We described behavior at a gross level and did not distinguish between specific treatment methods or rationales that would have enabled finer distinctions between doctors' behavior. We also reiterate that the variables in the analysis were generated by a single coder, with subsequent recoding used to demonstrate acceptable reliability.

An additional limitation imposed by these methods is sample size. Though the sample was quite large for an observational study, the data set was not as large as would typically be used for some of our quantitative analyses. This raises the possibility that our results may be due to sampling error or may not generalize to other emergency departments due to peculiarities of the specific settings in which our observations took place. Limited sample size may also have reduced statistical power to identify an effect of shift difficulty in the quantitative analysis.

Range restriction may also have limited our ability to detect potential effects of shift difficulty. The doctors we observed self-limited the number of patients seen at any one time, indirectly imposing an upper limit on total shift difficulty. Another limitation related to shift difficulty is the relationship between shift difficulty and the care of any individual patient. Individual patients' length of stay rarely spanned an entire shift; physicians' workload varied at any given moment and patients whose stays in the ED did not overlap would likely not affect one another. A shift-wide measure may not adequately capture the doctors' workload at the time that any individual patient was seen. However, it would be difficult to calculate a more patient-specific measure of shift difficulty given the continuous changes in a doctor's overall workload during the course of any individual patient's stay in the ED.

Finally, we assume that physicians' self-directed behavior is in fact ecologically rational with respect to outcomes. In other words, we assume that the behaviors we observed not only helped the doctors cope with changing situational demands from a cognitive or workload management perspective, but were also related to good patient outcomes. However, we have no way to actually validate whether the contextual

adjustments we observed allowed doctors to deliver quality care.

Future Research

This study raises several additional interesting questions for future research, some of which are described here. First, this study worked within a single domain (emergency medicine) in a single part of the United States. Future work should investigate whether goal establishment and goal enactment constructs generalize across domains and, if so, whether a similar pattern of environmental influences apply. Next, this study has identified the effects of nonmedical patient characteristics and the effect of the work system on decision making. These new insights should form the basis upon which new experimental tasks can be designed to capture a more complete range of behavior and allow for causality claims regarding the relationships observed here. On the observational side, research is also needed to establish the relationship between the processes identified here and outcome. Finally, we analyzed our sample in the aggregate without exploring variations due to differing experience among the physicians. Future work will generate a larger sample of attending physicians so that we can analyze the resident and attending physicians separately. Such an analysis would allow us to generate hypotheses regarding the acquisition of context sensitivity to provide new insight into the development of medical expertise over time.

Conclusions

This study has identified two key components of decision making (goal establishment and goal enactment) in U.S. emergency departments. Contextual influences on both processes highlight ways in which nominally similar problems may nonetheless be unique. Such effects also underscore nontechnical contributors to the competent practice of medicine and call into question the decontextualized patient as the unit of analysis in psychological studies of medical reasoning research, which may obscure other important influences on decision-making behavior. Further, differences across work setting underscore the difficulty in assessing quality of care based on outcome measures such as readmission alone. By making the relevant contextual influences on physician behavior (and the resulting adaptations on the part of the physicians) explicit, we hope to expand the experimental research on medical decision making to capture a more complete set of key variables and inform the development of outcome metrics that acknowledge and better accommodate local variability in the care setting.

ORCID iD

Frank Eric Robinson https://orcid.org/0000-0002-5723-5658

SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

REFERENCES

Arora, S., Sevdalis, N., Nestel, D., Woloshynowych, M., Darzi, A., & Kneebone, R. (2010). The impact of stress on surgical performance: A systematic review of the literature. *Surgery*, *147*(3), 318–330.

Atkins, S., & Ersser, S. J. (2008). Clinical reasoning and patient-centered care. In J. Higgs, M. Jones, S. Loftus, & N. Christensen (Eds.), *Clinical reasoning in the health professions* (pp. 68–77). Elsevier.

Benner, P. (1982). From novice to expert. *The American Journal of Nursing*, *82*(3), 402–407.

Bliese, P. D. (2002). Multilevel random coefficient modeling in organizational research: Examples using SAS and S-PLUS. In F. Drazgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 401–445). Jossey-Bass, Inc.

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.

Brooks, L. R., Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General*, *120*(3), 278–287.

Carayon, P., Alyousef, B., & Xie, A. (2012). Human factors and ergonomics in health care. In G. Salvendi (Ed.), *Handbook of human factors and ergonomics* (4th ed.). John Wiley & Sons.

Croker, A., Loftus, S., & Higgs, J. (2008). Multidisciplinary clinical decision making. In J. Higgs, M. Jones, S. Loftus, & N. Christensen (Eds.), *Clinical reasoning in the health professions* (pp. 291–298). Elsevier.

Croskerry, P. (2013). From mindless to mindful practice — cognitive bias and clinical decision making. *New England Journal of Medicine*, *368*(26), 2445–2448.

Cummings, E. M. (1980). Caregiver stability and day care. *Developmental Psychology*, *16*(1), 31–37.

Currey, J., & Botti, M. (2003). Naturalistic decision making: A model to overcome methodological challenges in the study of critical care nurses' decision making about patients' hemodynamic status. *American Journal of Critical Care*, *12*(3), 206–211.

Donner-Banzhoff, N., Seidel, J., Sikeler, A. M., Bösner, S., Vogelmeier, M., Westram, A., Feufel, M., Gaissmaier, W., Wegwarth, O., & Gigerenzer, G. (2017). The phenomenology

of the diagnostic process: A primary care-based survey. *Medical Decision Making : An International Journal of the Society for Medical Decision Making*, *37*(1), 27–34.
Dreyfus, H. L., & Dreyfus, S. E. (2005). Expertise in real world contexts. *Organization Studies*, *26*(5), 779–792.
Elstein, A. S. (2009). Thinking about diagnostic thinking: A 30-year perspective. *Advances in Health Sciences Education*, *14*(S1), 7–18.
Engestrom, Y. (1993). Developmental studies of work as a testbench of activity theory: The case of primary care medical practice. In S. Chaiklin & J. Lave (Eds.), *Understanding practice: Perspectives on activity and context* (pp. 64–103). Cambridge University Press.
Feufel, M. (2009). *Bounded rationality in the emergency department* [Unpublished doctoral dissertation]. Wright State University.
Feufel, M. A. (2018). How to uncover sources of unwarranted practice variation: A case study in emergency medicine. *Qualitative Health Research*, *28*(9), 1486–1498.
Feufel, M. A., & Flach, J. M. (2019). Medical education should teach heuristics rather than train them away. *Medical Education*, *53*(4), 334–344.
Foy, R., Hempel, S., Rubenstein, L., Suttorp, M., Seelig, M., Shanman, R., & Shekelle, P. G. (2010). Meta-analysis: Effect of interactive communication between collaborating primary care physicians and specialists. *Annals of Internal Medicine*, *152*(4), 247–258.
Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science*, *3*(1), 20–29.
Gilboy, N., Tanabe, P., Travers, D., & Rosenau, A. (2011). Emergency Severity Index (ESI): A triage tool for emergency department care. Implementation Handbook Version 4. Agency for Healthcare Research and Quality document #12-0014. Retrieved August 28, 2019, from https://www.ahrq.gov/sites/default/files/wysiwyg/professionals/systems/hospital/esi/esihandbk.pdf
Gonzales, K. (2010). Medication administration errors and the pediatric population: A systematic search of the literature. *Journal of Pediatric Nursing*, *25*(6), 555–565.
Greenhalgh, T., Howick, J., Maskrey, N., & Evidence Based Medicine Renaissance Group. (2014). Evidence based medicine: A movement in crisis? *BMJ*, *348*, g3725.
Henkel, R. J., & Maryland, P. A. (2015). The risks and rewards of value-based reimbursement. *Frontiers of Health Services Management*, *32*(2), 3–16.
Ho, N., Sadler, G. G., Hoffmann, L. C., Zemlicka, K., Lyons, J., Fergueson, W., Richardson, C., Cacanindin, A., Cals, S., & Wilkins, M. (2017). A longitudinal field study of Auto-GCAS acceptance and trust: First-year results and implications. *Journal of Cognitive Engineering and Decision Making*, *11*(3), 239–251.
Hutchins, E. (2014). The cultural ecosystem of human cognition. *Philosophical Psychology*, *27*(1), 34–49.
Johnson, J., Miller, S., & Horowitz, S. (2008). Systems-based practice: Improving the safety and quality of patient care by recognizing and improving the systems in which we work. In K. Henriksen, J. Battles, M. Keyes, & M. Grady (Eds.), *Advances in patient safety: New directions and alternative approaches (Vol. 2: culture and redesign)*. Agency for Healthcare Research and Quality.
Klein, G. (2007a). Flexecution as a paradigm for replanning, part 1. *IEEE Intelligent Systems*, *22*(5), 79–83.
Klein, G. (2007b). Flexecution, part 2: Understanding and supporting flexible execution. *IEEE Intelligent Systems*, *22*(6), 108–112.
Kliff, S., & Keating, D. (2013, May 8). One hospital charges $8,000 – another, $38,000. *The Washington Post*. https://www.washingtonpost.com/news/wonk/wp/2013/05/08/one-hospital-charges-8000-another-38000/?noredirect=on
Kulatunga-Moruzi, C., Brooks, L. R., & Norman, G. R. (2004). Using comprehensive feature lists to bias medical diagnosis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 563–572.
LaHuis, D. M., & Ferguson, M. W. (2009). The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods*, *12*(3), 418–435.
Lippa, K. D., Feufel, M. A., Robinson, F. E., & Shalin, V. L. (2017). Navigating the decision space: Shared medical decision making as distributed cognition. *Qualitative Health Research*, *27*(7), 1035–1048.
Martin, L., Haskard-Zolnierek, K., & DiMatteo, M. R. (2010). *Health behavior change and treatment adherence*. Oxford University Press.
Maude, J. (2014). Differential diagnosis: The key to reducing diagnosis error, measuring diagnosis and a mechanism to reduce healthcare costs. *Diagnosis*, *1*(1), 107–109.
Norman, G. R., & Eva, K. W. (2010). Diagnostic error and clinical reasoning. *Medical Education*, *44*(1), 94–100.
Reyna, V. F., & Lloyd, F. J. (2006). Physician decision making and cardiac risk: Effects of knowledge, risk perception, risk tolerance, and fuzzy processing. *Journal of Experimental Psychology: Applied*, *12*(3), 179–195.
Roter, D., & Hall, J. (2006). *Doctors talking with patients - patients talking with doctors*. Praeger.
Shalin, V. L., & Bertram, D. A. (1996). Functions of expertise in a medical intensive care unit. *Journal of Experimental & Theoretical Artificial Intelligence*, *8*(3-4), 209–227.
Shalin, V. L., Geddes, N. D., Bertram, D. L., Szczepkowski, M., & DuBois, D. (1997). Expertise in dynamic, physical task domains. In P. Feltovich, K. Ford, & R. Hoffman (Eds.), *Expertise in context: Human and machine*. AAAI Press.
Singh, M. (2018, December 3). Kids with concussions can phase in exercise, screen time Sooner than before. *NPR*. https://www.npr.org/sections/health-shots/2018/12/03/672002830/kids-with-concussions-can-phase-in-exercise-screen-time-sooner-than-before
Smith, M., Higgs, J., & Ellis, E. (2008). Factors influencing clinical decision making. In J. Higgs, M. Jones, S. Loftus, & N. Christensen (Eds.), *Clinical reasoning in the health professions* (3rd ed., pp. 89–100). Elsevier.
Timmermans, S., & Mauck, A. (2005). The promises and pitfalls of evidence-based medicine. *Health Affairs*, *24*(1), 18–28.
Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart. *Current Directions in Psychological Science*, *16*(3), 167–171.
van der Veen, W., van den Bemt, P. M. L. A., Wouters, H., Bates, D. W., Twisk, J. W. R., de Gier, J. J., Taxis, K., Duyvendak, M., Luttikhuis, K. O., Ros, J. J. W., Vasbinder, E. C., Atrafi, M., Brasse, B., Mangelaars, I., & BCMA Study Group. (2018). Association between workarounds and medication administration errors in bar-code-assisted medication administration in hospitals. *Journal of the American Medical Informatics Association*, *25*(4), 385–392.
Walsh, T., & Beatty, P. C. W. (2002). Human factors error and patient monitoring. *Physiological Measurement*, *23*(3), R111–R132.
Ward, P., Gore, J., Hutton, R., Conway, G. E., & Hoffman, R. R. (2018). Adaptive skill as the *condition sine qua non* of expertise. *Journal of Applied Research in Memory and Cognition*, *7*(1), 35–50.
Ward, P., Hutton, R., Hoffman, R., Gore, J., Anderson, T., & Leggatt, A. (2016). *Developing skilled adaptive performance: A scoping study: Final technical report*. BAE Systems.
Weiss, D. J., & Shanteau, J. (2014). Who's the best? A relativistic view of expertise. *Applied Cognitive Psychology*, *28*(4), 447–457.

Frank Eric Robinson is a research psychologist at the Naval Medical Research Unit Dayton after earning his PhD from Wright State University in 2017. He studies expertise, environmental influences on cognition, and the relationship between deliberate and automatic processes in behavior. He has studied shared decision making between patients and providers, the effects of medical records systems on work practice, ways to evaluate surgical performance, and team coordination processes among USAF critical care aircrew.

Markus A. Feufel heads the Division of Ergonomics in the Department of Psychology and Ergonomics at Technische Universität Berlin. He is an associated scientist in the Harding Center for Risk Literacy and the Institute of Medical Sociology and Rehabilitation Science at Charité-Universitätsmedizin Berlin. His research investigates how work systems adapt to changing demands and how to empower individuals in these systems, including patients and providers in health systems, to help shape adaptation effectively and actively.

Valerie L. Shalin is a professor of psychology at Wright State University, having earned her PhD in learning, developmental, and cognitive psychology from the University of Pittsburgh in 1987. A recipient of the 2016 Human Factors Prize, she conducts research on planning and communication processes in coordinated work and corresponding workplace technology for space exploration, disaster response, and manual labor. Her work on these issues in medicine and surgery spans more than two decades.

Debra Steele-Johnson is a professor of psychology at Wright State University. She received her PhD in industrial/organizational psychology from the University of Minnesota in 1988. Her research focuses on how people acquire and perform complex skills in individual and team contexts and factors that affect those processes, including task features, training, leadership, motivation, and personal characteristics.

Brian Springer, MD, is an associate professor at the Wright State University Department of Emergency Medicine in Dayton, OH. He has been full-time faculty with the Department of Emergency Medicine since 2002 and has served as director of the WSU Division of Tactical Emergency Medicine since its inception in 2009. Dr. Springer is an attending physician at the Kettering Medical Center Emergency Department.

Effects of Workload and Workload Transitions on Attention Allocation in a Dual-Task Environment: Evidence From Eye Tracking Metrics

Nadine Marie Moacdieh, American University of Beirut, Lebanon, **Shannon P. Devlin**, University of Virginia, Charlottesville, VA, USA, **Hussein Jundi**, American University of Beirut, Lebanon, and **Sara Lu Riggs**, University of Virginia, Charlottesville, VA, USA

High mental workload, in addition to changes in workload, can negatively affect operators, but it is not clear how sudden versus gradual workload transitions influence performance and visual attention allocation. This knowledge is important as sudden shifts in workload are common in multitasking domains. The objective of this study was to investigate, using performance and eye tracking metrics, how constant versus variable levels of workload affect operators in the context of a dual-task paradigm. An unmanned aerial vehicle command and control simulation varied task load between low, high, gradually transitioning from low to high, and suddenly transitioning from low to high. Performance on a primary and secondary task and several eye tracking measures were calculated. There was no significant difference between sudden and gradual workload transitions in terms of performance or attention allocation overall; however, both sudden and gradual workload transitions changed participants' strategy in dealing with the primary and secondary task as compared to low/high workload. Also, eye tracking metrics that are not frequently used, such as transition rate and stationary entropy, provided more insight into performance differences. These metrics can potentially be used to better understand operators' strategies and could form the basis of an adaptive display.

Keywords: workload, topics, eye movements, attention

Address correspondence to Nadine Marie Moacdieh, Department of Industrial Engineering and Management, American University of Beirut, Beirut, Lebanon, nm102@aub.edu.lb

Journal of Cognitive Engineering and Decision Making
2020, Volume 14, Number 2, June 2020, pp. 132–151
DOI:10.1177/1555343419892184

INTRODUCTION

High mental workload can cause significant performance decrements and breakdowns in attention allocation (Rouse et al., 1993). Preventing problems related to high workload is critical in complex, safety-sensitive domains, such as aviation, process control, driving, and medicine (e.g., Dixon et al., 2005, Van Benthem et al., 2015), where delays and errors due to high workload can have life-threatening consequences. The 1978 United Airlines DC-8 airplane crash in Portland, OR, for example, was linked to pilots' high workload as they were trying to address issues with the landing gear (National Transportation Safety Board, 1978).

Workload is a multidimensional construct for which there are several definitions in the literature; we have adopted Wickens' (1992) definition of workload as the gap between one's attentional resources and the cognitive demands placed on the user. One common way to influence the cognitive demands of a person is by manipulating their task load (Hancock et al., 1995). Task load can be defined as the number of items that one has to attend to in order to successfully complete a task (e.g., Veltman & Gaillard, 1996). While research on high workload has fully established its detrimental effects on performance (e.g., Brookings et al., 1996; Dixon & Wickens, 2006; Matthews et al., 2015), studies tend to assess the effects of workload in discrete and separate time intervals. However, this does not always typify the work of operators in real-world situations, where workload can fluctuate over time. This has led to research on what has been termed workload transitions (Huey & Wickens, 1993; Prytz & Scerbo, 2015), workload history (Cox-Fuenzalida, 2007),

or hysteresis (Morgan & Hancock, 2011). The literature has shown that workload transitions negatively affect operator performance, suggesting the concern about fluctuating workload levels is well founded (Cox-Fuenzalida et al., 2006; 2007; Cumming & Croft, 1973; Goldberg & Stewart, 1980).

However, despite these efforts to study the effects of workload transitions, there are still three major gaps in the literature on this topic. The first gap is the lack of emphasis on the difference between gradual and sudden workload changes, as opposed to just analyzing the difference between low-to-high versus high-to-low workload transitions (e.g., Cumming & Croft, 1973; Goldberg & Stewart, 1980). An air traffic control operator, for example, may have to deal with a sudden increase in the number of airplanes to attend to; this could affect the operator's performance differently than if the number of planes increased gradually over time. In other words, it could be that the element of surprise in sudden workload transitions could lead to additional decrements in performance (Kochan et al., 2004; Wickens, 2001).

The second gap is the absence of workload transition studies occurring in a more realistic environment, especially ones that involve multitasking. For instance, Cox-Fuenzalida (2007) investigated the effects of workload transitions by presenting participants with various number strings and asking them to identify particular sequences of digits. The study showed that, contrary to previous research (e.g., Matthews, 1986), a sudden increase in workload did lead to performance decrements. While such studies are valuable for understanding workload transitions, they do not necessarily generalize to more complex environments, such as air traffic control operations or military environments, where operators may have more than one task and several different areas of the screen to attend to.

The third and final gap is the type of approach used to measure the effects of workload transitions, which currently do not capture changes in visual attention allocation in detail. Eye tracking technology is a promising tool in this regard. Several eye tracking metrics have been shown to be sensitive to differences between low and high workload (see Coral, 2016 for a review), but eye tracking has not been used to date to examine the changes in attention allocation that result from sudden and gradual workload transitions in comparison to constant low and high workload.

Objectives

The aim of this research study was to analyze how performance and attention allocation, as evidenced by eye movements, are affected by gradual and sudden workload transitions as compared to constant low or high workload in a multitasking environment. Our expectations were that sudden workload transitions would result in more performance detriments compared to gradual ones, based on the reported detrimental effects of sudden workload transitions (e.g., Cox-Fuenzalida, 2007; Kochan et al., 2004). It was also expected that performance on both the sudden and gradual conditions would be midway between low workload (best) and high workload (worst). The eye tracking measures, though, were the main focus of this study, and it was expected that these measures would help provide insight into how the performance effects came about. More specific hypotheses are provided at the end of the "Background" section, after the eye tracking metrics are defined.

To this end, a realistic simulation environment was used, with variations in task load used as a means of modulating mental workload. The focus in this study was on the types of tasks that require monitoring several spatially dispersed areas of a screen and searching for certain targets, such as what could be found in military applications, air traffic control operations, and security system monitoring. The selected application domain was unmanned aerial vehicle (UAV) operations, an example of a multitasking domain where workload fluctuates.

BACKGROUND

Workload and Workload Transitions

Several theories have attempted to explain what happens during workload transitions. However, the well-known theories that have been posited—Cumming and Croft's (1973) error acceptance theory, Goldberg and Stewart's (1980) short-term memory theory, Matthews' (1986) task overworking theory, and

Cox-Fuenzalida's (2007) adaptation model—focus more on why high to low workload transitions tend to be more detrimental than the inverse. There is no satisfactory explanation or focus in these studies on the differences between sudden and gradual changes in workload levels, least of all at the level of attention allocation.

Approaches that have usually been adopted to examine workload include subjective ratings, performance measures, and physiological measures (Cain, 2007). Subjective ratings include the widely used NASA-Task Load Index (TLX; Hart & Staveland, 1988) and the Subjective Workload Assessment Technique (Reid & Nygren, 1988). The challenge with subjective measures is that they cannot provide detailed information regarding attention allocation. Likewise, while performance measures, such as response time and error rate (e.g., Bliss & Dunn, 2000; Boyer et al., 2015; Dixon & Wickens, 2006), provide a good impression of the effects of low and high workload, they are also not well suited for tracing changes in attention allocation. Physiological measures, on the other hand, are better suited for that purpose. These include measures such as electroencephalography (EEG; Berka et al., 2007), electrocardiogram recordings (ECG; e.g., Solovey et al., 2014), heart rate variability (e.g., Hoover et al., 2012; Schulz et al., 2011), galvanic skin response (GSR; e.g., Solovey et al., 2014), and eye tracking—the focus of this study.

Eye Tracking

Eye tracking is a technique used to trace where a user is looking on a display, typically using infrared light (Poole & Ball, 2005). Eye tracking can be used to track approximate gaze location, which in turn indicates overt attention allocation or the shifting of the eyes toward a stimulus (Bergen & Julesz, 1983; Treisman & Gelade, 1980). This is in contrast to covert attention, which relates to mentally focusing on a certain stimulus. A person could be looking at an item and thinking about something completely different, but the eye tracker cannot recognize this discrepancy and would only indicate where the person is looking. In addition, it is important to note that eye tracking output only provides information about foveal vision—that is, the high acuity central vision that is used to visually process items in detail—but less about peripheral vision (Rosenholtz, 2016) or the useful field of view (Wolfe et al., 2017), which are both relevant to visual perception. This means that there could be information that participants are obtaining that is not directly reflected in the eye tracking metrics. This limitation of eye trackers, however, has not prevented the approach being widely used in human factors research as a means to understand visual attention allocation (Duchowski, 2007).

In contrast to other measures of assessing workload, such as EEG, ECG, and GSR, eye tracking is usually nonobtrusive, meaning that the user is not physically tethered or attached to anything. Moreover, eye tracking can be used to assess a person's situation awareness (SA), which is defined as people's perception of the items around them, their understanding of these items, and their estimate of their state in the near future (Endsley, 1995). However, assessing SA is a challenge because current SA measures that have high reliability and validity require frequently interrupting the participant to ask questions/probes, for example, Situation Awareness Rating Technique (Taylor, 1990) and Situation Awareness Global Assessment Techniques (Endsley, 1988). Eye tracking, on the other hand, would not disturb the operator during a task.

The output from an eye tracker is a series of gaze points that allow researchers to assess when and for how long users were looking at screen elements. The coordinates are used to determine eye *fixations*, or spatially stable gaze points during which visual processing takes place, and *saccades,* which are ballistic movements of the eye between two fixations (Holmqvist et al., 2011).

The most frequently used eye tracking measures for workload evaluation are pupillometry metrics such as pupil diameter (Hampson et al., 2010) and eye blink frequency and duration (Hwang et al., 2008; Veltman & Gaillard, 1996). A full discussion and meta-analysis of these metrics can be seen in Coral (2016). It has been found that these measures are positively correlated with workload; however, such measures

are also sensitive to other factors, such as the amount of light (Monfort et al., 2016). Avoiding this problem would require keeping the amount of light constant at all times, incorporating ambient light into the estimate of workload, or incorporating some other measure of workload, such as EEG (e.g., Rozado & Dunser, 2015).

Alternatives to these pupillometry metrics do exist and were the focus of this research study. The selected eye tracking metrics, together with their definitions and our hypotheses regarding each metric, are summarized in Table 1. Mean fixation duration and saccade amplitude have both been used in workload research (De Rivecourt et al., 2008; Di Stasi et al., 2013), as has the nearest neighbor index (NNI; Di Nocera et al., 2007). NNI has been used to determine how spread out or concentrated fixations are and has been shown to be sensitive to workload in various domains that include aviation (Di Nocera et al., 2007) and health care (Moacdieh & Sarter, 2015a). Other measures that have been shown to be useful are stationary and transition entropy (Krejtz et al., 2014). These entropy measures provide an estimate of fixation sequence randomness and have been used in previous studies to estimate workload (e.g., Monfort et al., 2016).

The metrics in Table 1 have been categorized by the *spread* (where are users looking?), *directness* (how efficiently are users scanning?), and *duration* (how long are users looking at a certain area?) of eye gaze points, as classified in Moacdieh and Sarter (2015b). All of these metrics have been used to a much lesser extent than pupillometry metrics, but they offer two major advantages over the latter. First, the metrics in Table 1 are not as sensitive to light as the pupillometry metrics (controlling for light variation can attenuate this problem, although that could be hard to do in a realistic environment). Second, the metrics provide a more reliable and comprehensive assessment of how workload affects scan patterns than just measuring pupil size or blink frequency. The metrics have been used to provide useful insight into the effects of display clutter (Moacdieh & Sarter, 2015a); however, spread metrics have rarely been explored in the context of workload (e.g., Rantanen & Goldberg, 1999) and only a small selection of directness metrics, such as mean saccade amplitude (e.g., Savage et al., 2013), have been explored in this context. In particular, directness metrics, which focus on the efficiency of users' scan patterns, have the potential to reveal more insights about a person's use of a display in divided attention scenarios such as the ones in this study (as detailed in the "Methods" section). Since participants will be asked to deal with a primary task and a number of secondary tasks, with each task located in a different area of the screen, larger saccade amplitudes and more transitioning between areas of the screen would suggest uncertainty about the location of the target or about the task to focus on. For example, questions that directness measures can answer include: *How frequently did operators transition between multiple areas of interest? Did they use the shortest path to reach their goal? Was there a lot of inefficient back-and-forth scanning*? By answering such questions, directness metrics can provide insight into the efficiency of users' scan patterns. Efficiency of eye movement scanning is related to how much screen distance the user's eyes had to travel in order to complete the task, with more efficient users being ones who cover less distance.

METHODS

Participants

Twenty-one students participated in this study (13 men and 8 women; mean age = 20.9, *SD* = 1.5). Participants had self-reported normal or corrected-to-normal vision. Participants were compensated $10/hr for their participation. Participants gave informed consent and the study was conducted in accordance with the tenets of the Declaration of Helsinki. The study was approved by the Clemson University Institutional Review Board (IRB2015-217).

Experimental Setup

The simulation was developed using the Unity game development platform and was based on the "Vigilant Spirit Control Station" used by the Air Force to develop interfaces to control multiple UAVs (Feitshans et al., 2008). The simulation was displayed on the full screen of a desktop computer with a 32″ monitor (2,560 × 1,600 screen resolution). A Fovio eye

TABLE 1: Eye Tracking Metrics Investigated as Part of This Study, Together with the Definition of Each and Our Hypotheses for How They Will Change in the Experiment

Metric	Definition and Calculation	Hypotheses
Spread Metrics		
Convex hull area (pixels^2)	The minimum convex area which contains the fixation points (Goldberg & Kotval, 1999). This is calculated using the Matlab function convHull, with the X and Y positions of the fixation points as input. The maximum area of the screen is 2,560 × 1,600 = 4.096 ×10^6 pixels^2 A larger convex hull area indicates more spread of gaze points and larger cognitive load as the user attempts to sample all the information available within the display (Di Nocera et al., 2007)	H1: As performance deteriorates, these metrics will exhibit increased spread as participants distribute their visual attention to many and wide-ranging areas of the display
Spatial density	The number of grid cells containing gaze points divided by the total number of cells (Goldberg & Kotval, 1999). A 20 × 20 evenly divided grid (128 × 80 pixels per cell) was created to cover the full screen dimensions. Similar to convex hull area, a higher spatial density would indicate a larger dispersion of attention	
Stationary entropy	Stationary entropy indicates how equally distributed a person's attention is, with larger values indicating more evenly spread attention across areas of interest and lower values indicating more narrowed attention (Krejtz et al., 2014). Stationary entropy is calculated using the following equation $H_s = -\sum_{i \in \text{AOIs}} \pi_i log \pi_i$ where π_i represents the long-run fraction of time the chain spends in the *i*th state. For our purposes, it represents the long-run fraction of time a participant spent in the *i*th AOI (the AOIs are as defined in Figure 1). Assuming the properties of a first-level Markov chain hold, π_i can be determined by solving the system of equations $\boldsymbol{\pi P} = \boldsymbol{\pi}$, where $\boldsymbol{\pi}$ is stationary distribution of the chain (i.e., every π_i in the chain) and $\boldsymbol{P}$ is a one-step transition matrix of the chain	
Directness Metrics		
Mean saccade amplitude (pixels)	The average amplitude of saccades. Higher mean saccade amplitude indicates lower scanning efficiency	

(Continued)

TABLE 1 (Continued)

Metric	Definition and Calculation	Hypotheses
Scanpath length per second (pixels/s)	The sum of all the saccade lengths divided by the total time. Similar to mean saccade amplitude, a larger scanpath length indicates less efficiency	H2: As performance deteriorates, these metrics will exhibit less directness and efficiency as participants switch their visual attention more frequently and randomly among areas of the display
Backtrack rate (s^{-1})	A backtrack is defined as an angle between two saccades that is greater than 90° (Goldberg & Kotval, 1999), indicating a change in direction. A higher backtrack rate indicates lower efficiency	
Transition rate (s^{-1})	The rate of transitions between equal grid cells (Goldberg & Kotval, 1999). A higher rate of transitions indicates lower efficiency. The same grid cells used for spatial density were used here	
Transition entropy	The transition entropy represents the randomness and complexity of a person's eye movements, with higher values indicating more randomness and lower efficiency (Krejtz et al., 2014). The transition entropy was calculated based on the following formula: $H_t = -\sum_{i \in AOIs} \pi_i \sum_{j \in AOIs} p_{ij} \log p_{ij}$ where π_i is calculated as for stationary entropy, and p_{ij} is the probability of transitioning from state *i* to state *j* in one unit of time. Assuming the assumption for a first-order Markov chain holds, this was calculated by counting the number of transitions from *i* to *j* and then dividing by the total number of transitions from *i*. This was done for each pairing of AOIs (the AOIs are as defined in Figure 1)	
Duration Metric		
Mean fixation duration (ms)	A lower mean fixation duration suggests that the user is quickly moving from one focus to the next	H3: As performance deteriorates, there will be increased mean fixation duration as participants struggle to discriminate information from the display

Note that any mention of attention here refers to overt attention allocation.

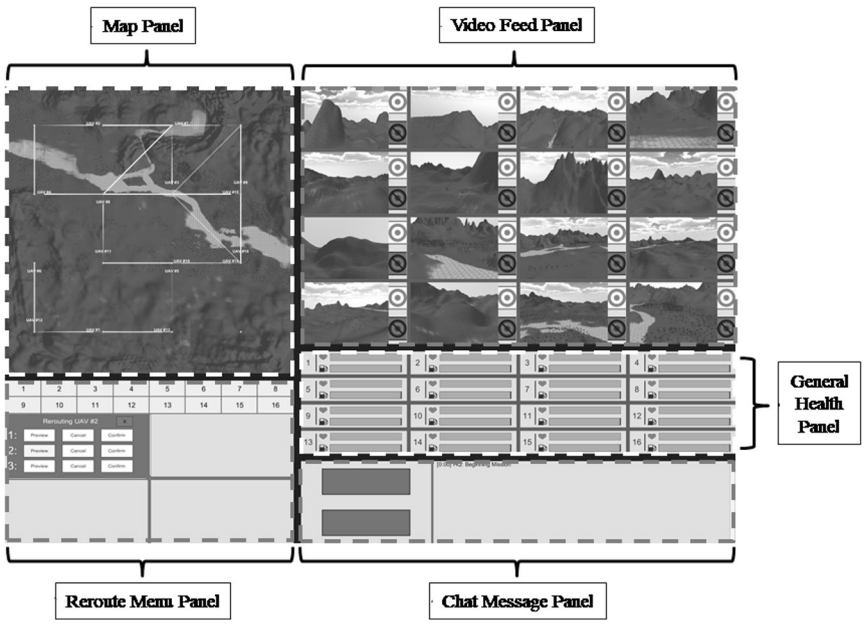


Figure 1. Screenshot of the UAV simulation with labeled panels. Each of these panels also constituted the AOIs for the calculation of the entropy measures.

tracker, a desktop-mounted eye tracker with a sampling rate of 60 Hz, was used to collect eye tracking data. The eye tracker was placed right below the monitor. Participants sat 28″–31″ from the monitor and used a standard mouse to input responses. The average degree of error for this eye tracker is 0.78 degrees (*SD* = .59; (Eyetracking, 2011).

UAV Control Simulation and Tasks

Participants were responsible for simultaneously controlling and managing 16 UAVs under four different task load conditions that mapped to four different scenarios (see "Task Load Modulation" section). Each scenario was 15 min in duration. This time frame was selected as it would be long enough to include multiple transitions from low and high workload, while not being too long where significant vigilance decrements could occur (typically considered around 15 min, although several factors can play a role; Teichner, 1974). Figure 1 depicts the simulation interface with task-specific areas of interest, referred to as panels, in dotted lines. The four tasks included one primary task (target detection) and three secondary tasks (reroute, fuel leak, and chat message task). For all four scenarios, the rate at which the primary task occurred was varied (see "Task Load Modulation" section) and one secondary task occurred every 20 s.

Target detection task (primary task). Target detection was the primary task and participants were instructed that this task had the highest priority. At the most fundamental level, the primary task was a visual search task for targets (transparent cubes on the video feed panel; Figure 2). There were up to 16 UAVs (i.e., 16 video feeds) at any given time. When a UAV approached a waypoint, the corresponding UAV video feed would become highlighted. When a UAV video feed was highlighted and a target was present, the participants were instructed to press the "target" button to indicate a target was present. Otherwise they were instructed to leave the default "no target" button selected. UAV video feeds were active for 10 s and targets could appear any time during this period. Of the UAV feeds that were active, a target was present

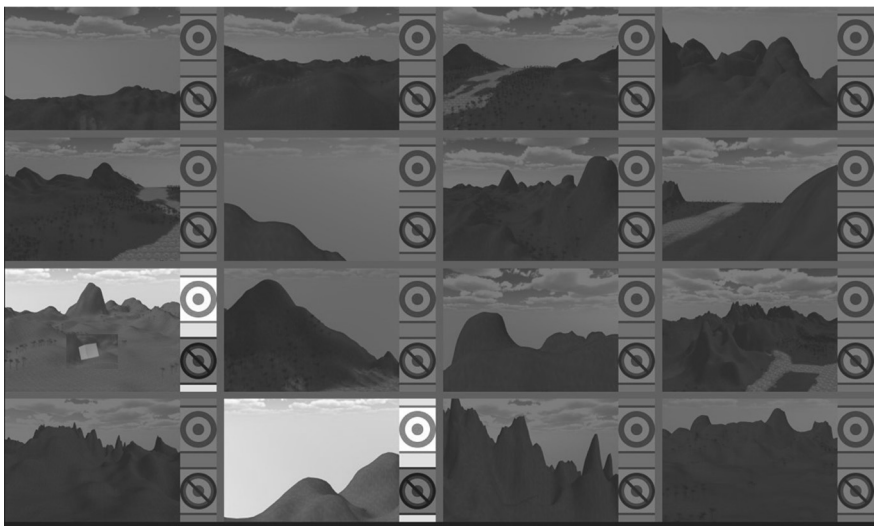


Figure 2. Screenshot of the video feed panel. UAVs 9 and 14 are highlighted (numbering from left to right, top to bottom). UAV 9 has a target (transparent cube, circled here for illustration purposes).

in 30% of those UAVs, on average. There could only be one target per UAV per 10 s interval. The specific UAVs that were concurrently highlighted at any point in time and the UAVs' arrival time to each waypoint were pseudorandomly set by the experimenter.

Reroute task (secondary task #1). The secondary tasks consisted of monitoring tasks for changes in the status of the UAVs. Participants were tasked to reroute a UAV when it entered a no-fly zone (i.e., red square on the map panel in Figure 3). To reroute a UAV, a participant clicked on the respective UAV's numbered square in the reroute menu panel (Figure 3). Participants had the option of selecting "Preview" to see three suggested routes, "Confirm" to reroute the UAV to a new suggested route, or "Cancel" to exit from the window to allow the UAV to continue on its original route. When a UAV was not rerouted in time and entered a no-fly zone, it became nonoperational for the remainder of the scenario (i.e., it could not participate in any of the target detection, reroute, or fuel leak tasks). Only three UAVs could enter a no-fly zone per scenario, meaning that even in the worst case scenario, if a participant failed to reroute all three UAVs, the number of UAVs available would still be within the range of active UAV needs in the high task load condition. The rerouting task occurred 18 times in each scenario.

Fuel leak task (secondary task #2). Participants were also tasked with maintaining the overall health of each UAV using the general health panel (Figure 4). Participants were asked to monitor for fuel leaks. When a fuel leak occurred, the color of the health status bar (top bar denoted with a heart) changed from green to yellow with a "FIX LEAK" warning. To stop a fuel leak, participants had 10 s to click on the yellow bar. If the leak was not stopped in time, the health status bar would change from yellow to orange and read "FATAL FUEL LEAK," meaning that the UAV could not participate in the fuel leak task again for that scenario. However, the UAV could still attend to the primary task, that is, the target detection task. A fuel leak occurred 14 times in each scenario.

Chat message task (secondary task #3). Participants were tasked with responding to chat messages by selecting between the two

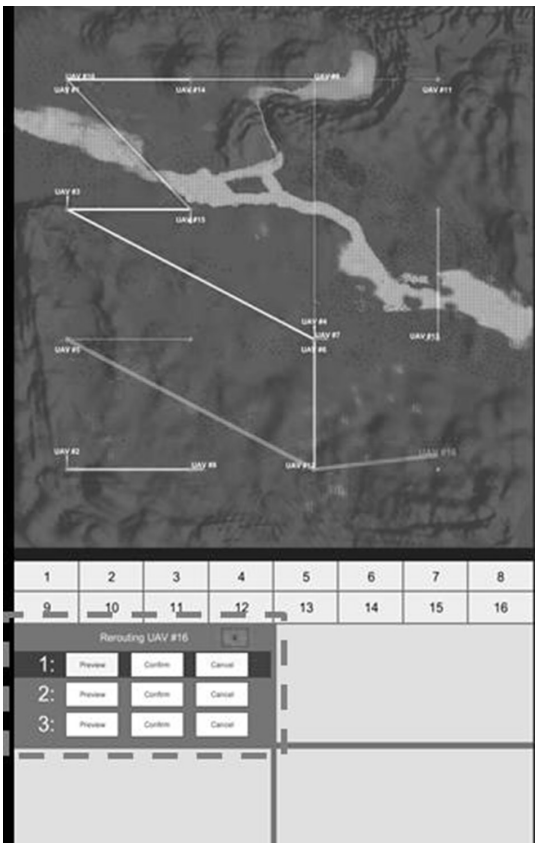


Figure 3. Screenshot of the map panel (top half) and reroute menu panel (bottom half). After clicking on the respective UAV's number (buttons numbered 1–16), a menu of route options was presented (dotted area with options 1–3 where "Preview," "Confirm," or "Cancel" could be selected).

options on the left-hand side of the chat message panel (Figure 5). Participants could respond to a chat message by clicking on one of two options until another message showed up. There were 19 chat messages in each scenario.

Response time and accuracy for the primary task were calculated for only the correct detection of targets. Cases where participants did not respond to the occurrence of a target were not considered in the calculation of response time. For all secondary tasks, response times were calculated from the onset of the event to when the participant responded. Accuracy was calculated as the percentage of correct responses (i.e., the percentage of correctly detected targets) within the time limit for each task.

TASK LOAD MODULATION

Task load (low, high, gradual, sudden) was the only independent variable and was varied fully within-subjects. Participants had to go through one scenario for each of the four task load conditions. Task load was manipulated by varying the number of active UAVs (the number of highlighted video feeds) in the target detection task. The four task load conditions include the following:

1. **Low task load.** There were three to five UAVs active for the entirety of the scenario. These values were determined and validated based on pilot testing data using both performance and NASA-TLX measures.
2. **High task load.** There were 13–16 UAVs active for the entirety of the scenario. These values were also determined based on pilot tests.
3. **Gradual task load.** The number of active UAVs increased gradually (Figure 6). The scenario started at low task load for 20 s, and one active UAV was added every 10 s until high task load was reached, that is, 13–16 active UAVs. The scenario would remain at high task load for 2 min, before returning to low task load. This low to high task load cycle repeated five times for this scenario.
4. **Sudden task load.** The number of active UAVs increased instantaneously (see Figure 6). One minute of low task load (3-5 UAVs) was followed by a jump to high task load (13-16 UAVs) for two minutes. This cycle repeated five times.

Note that the number of areas of the screen that participants had to monitor and attend to did not change based on task load condition because participants had to monitor all AOIs in order to complete all tasks.

Procedure

The experiment took place over two consecutive days around the same time of day (morning, afternoon, or evening). On the first day, participants signed a consent form and were briefed about the study's goals and expectations. Participants then completed a 5 min training session that included all tasks and task load conditions. By the end of the training session, participants had to demonstrate proficiency by having a minimum accuracy of 70% across all

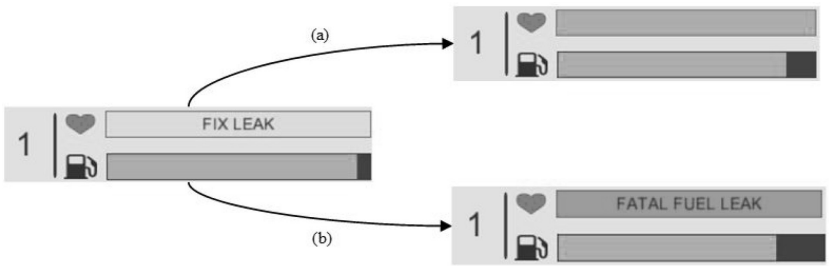


Figure 4. Screenshot of UAV 1's health status bar. Participants were tasked to press the yellow "FIX LEAK" button when a fuel leak occurred. (a) When a fuel leak was fixed in time, the yellow health status bar changed from yellow to green and the "FIX LEAK" warning disappeared. (b) When a fuel leak was not fixed in time, the yellow "FIX LEAK" changed to an orange "FATAL FUEL LEAK" warning.

training tasks, both primary and secondary. If they did not, they had another chance to complete the 5 min training session to achieve the desired proficiency. If they did not meet the proficiency requirements the second time, they were excused from the study. Only four participants had to repeat the training session, and all were successful the second time.

The eye tracker was calibrated using a 5-point grid. The calibration procedure was done at the start of each scenario. Participants completed two of the four scenarios on the first day and the other two on the second day, with the order counterbalanced across subjects. There was a 10 min break between scenarios. The entire study across 2 days lasted about 2 hr.

RESULTS

Results were analyzed using a one-way repeated measures ANOVA, with task load (four levels) as the variable of interest. Bonferroni corrections were applied for all post hoc tests. In all cases, Epsilon (ε) was calculated according to Greenhouse and Geisser (1959) and used to correct the one-way repeated measures ANOVA. For all graphs, error bars indicate the standard error of the mean and asterisks denote significant differences between conditions.

Figure 5. Screenshot of the chat message panel with response buttons on the left ("yes"/"no") and timestamped chat window on the right.

The initial gaze data were screened to meet data quality requirements as outlined in ISO/TS 15007-2:2014-09, which states that at most 15% data loss is acceptable for good quality data. Following this guideline, the eye tracking and corresponding performance data of five participants was not used in any of the analyses. The mean data loss of the included participants was 7.07%. The gaze points from the eye tracker were used to calculate fixations and saccades (the eye tracker automatically filters out blinks and any fixations outside the screen were discarded). Goldberg and Kotval (1999) fixation algorithm was applied: A cluster of gaze points was classified as a fixation if the points within the cluster were within 75 pixels of each other, and there was a minimum number of six gaze points within this fixation cluster. This made for a minimum fixation duration of approximately 100 ms. The first gaze point outside the 75-pixel limit was considered to be not part of the fixation; the gaze point just before would be the endpoint of the fixation. Any gaze points that were not part of fixations were assumed to be saccades. The calculated fixations were then used to calculate the eye tracking metrics described in Table 1.

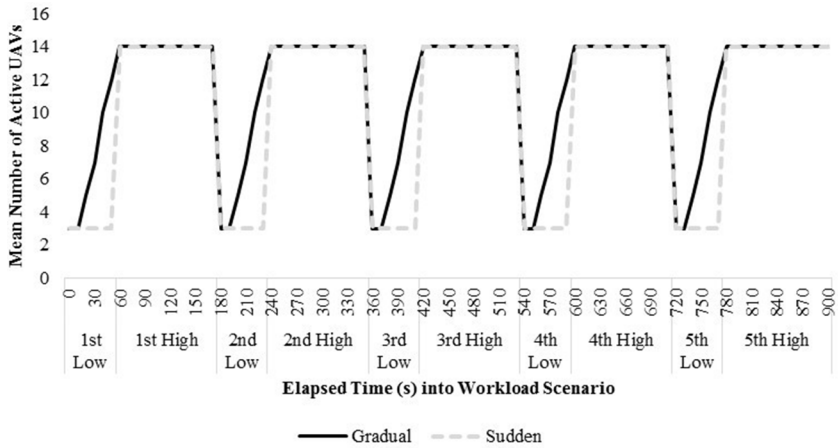


Figure 6. The number of active UAVs throughout the gradual and sudden task load scenarios.

Performance Results

Primary task. There was a significant effect of task load on response time (F[2.36,33.48] =161.51, $p < .001$, $\eta_p^2 = .95$, $\varepsilon = .94$), with means (standard error of the mean [*SE*]) equal to 2.37 s (*SE* = .04), 3.42 s (*SE* = .04), 2.86 s (*SE* = .04), and 2.86 s (*SE* = .03) in the low, high, gradual, and sudden conditions, respectively. Linear contrasts showed that the low task load scenario elicited significantly faster response times than all other scenarios (all $p < .001$). The high task load scenario had significantly slower participant response times than all other scenarios (all $p < .001$).

For primary task accuracy, there was a significant effect of task load on accuracy (F(2.61,39.25) = 101.42, $p < .001$, $\eta_p^2 = .87$, $\varepsilon = 1.00$), with means equal to 83.52% (*SE* = 1.20), 58.20% (*SE* = 1.50), 69.68% (*SE* = 1.90), and 72.09% (*SE* = 1.80) in the low, high, gradual, and sudden conditions, respectively. The low task load scenario led to significantly higher accuracy than all other scenarios (all $p < .001$). Participant accuracy in the high task load scenario was significantly lower than all other scenarios (all $p < .001$).

Secondary task. There was a significant effect of task load on secondary task response time (F(1.75,26.24) = 16.05, $p < .001$, $\eta_p^2 = .51$, $\varepsilon = .58$), with means of 5.09 s (*SE* = .28), 5.03 s (*SE* = .26), 3.88 s (*SE* = .10), and 3.70 s (*SE* = .15) in the low, high, gradual, and sudden conditions, respectively. Participant response time in the low task load scenario was significantly slower than the gradual and sudden task load scenarios (all $p < .001$). The high task load scenario was significantly slower than in the gradual and sudden task load scenarios (all $p < .001$).

There was a significant effect of task load on secondary task accuracy (F(2.02,30.40) = 7.15, $p = .003$, $\eta_p^2 = .32$, $\varepsilon = .67$), with means of 88.41% (*SE* = 2.80), 89.94% (*SE* = 2.40), 96.71% (*SE* = 1.00), and 96.08% (*SE* = 1.10) in the low, high, gradual, and sudden conditions, respectively. In the low task load scenarios, participants were significantly less accurate than in the gradual ($p = .006$) and sudden task load scenarios ($p = .003$). Participant accuracy in the high task load scenarios was also significantly lower than in the gradual ($p = .003$).

Eye Tracking Results

H1: spread metrics. The results of the spread metrics calculations can be seen in Figure 7. There was no significant difference of task load on convex hull area (Figure 7a). There was a significant effect of task load on spatial density (F(2.48,37.24) = 5.24, $p = .006$, $\eta_p^2 = .25$, $\varepsilon = .82$; Figure 7b). There were also significant pairwise

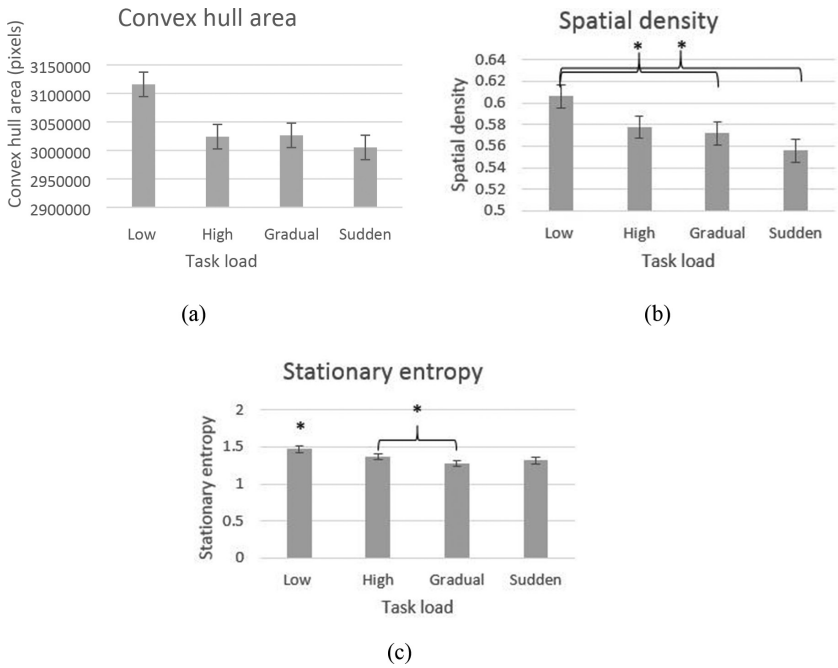


Figure 7. Spread metrics values for each task load condition: (a) convex hull area, (b) spatial density, and (c) stationary entropy.

differences between the low and gradual task load conditions (p = .004) and between low and sudden task load (p = .008). There was also a significant effect of task load on stationary entropy (F(2.21,33.17) = 9.49, p < .001, η_p^2 = .38, ε = .73; Figure 7c). Post hoc tests showed that low task load had a higher entropy than all other task load conditions, with p = .009, p = .001, and p = .007 for the high, gradual, and sudden conditions, respectively. High task load was also significantly higher than gradual (p = .014).

H2: directness metrics. For the directness metrics, which can be seen in Figure 8, there was a significant effect of task load type on mean saccade amplitude (F(2.38,35.76) = 13.81, p < .001, η_p^2 = .47, ε = .79; Figure 8a). Post hoc tests showed that there was a significant difference between low task load and all other conditions (all p < .001). For scanpath length per second, there was also a significant effect of task load (F(2.84,42.61) = 16.32, p < .001; η_p^2 = .52, ε = .94; Figure 8b). Post hoc tests showed that there was a significant difference between low task load and the other conditions (all p < .001). There was no significant effect of task load on backtrack rate (Figure 8c). There was a significant effect of task load on transition rate (F(2.65,39.75) = 4.92, p = .007, η_p^2 = .24, ε = .88; Figure 8d). There was a significant pairwise difference between the low and gradual (p = .029) and low and sudden (p = .005) conditions. For the transition entropy, there was an effect of task load (F(2.16,32.48) = 15.83, p < .001, η_p^2 = .51, ε = .72; Figure 8e). The low task load condition was significantly higher than all other conditions (p < .001, p < .001, and p = .001 for high, gradual, and sudden task loads, respectively). High task load was also significantly higher than the gradual task load condition (p = .007).

H3: duration metric. Finally, for the duration metric, there was no significant effect of task load on mean fixation duration (Figure 9). Table 2 provides a summary of the eye tracking results obtained in this study.

DISCUSSION

The aim of this study was to analyze how performance and attention allocation, as evidenced

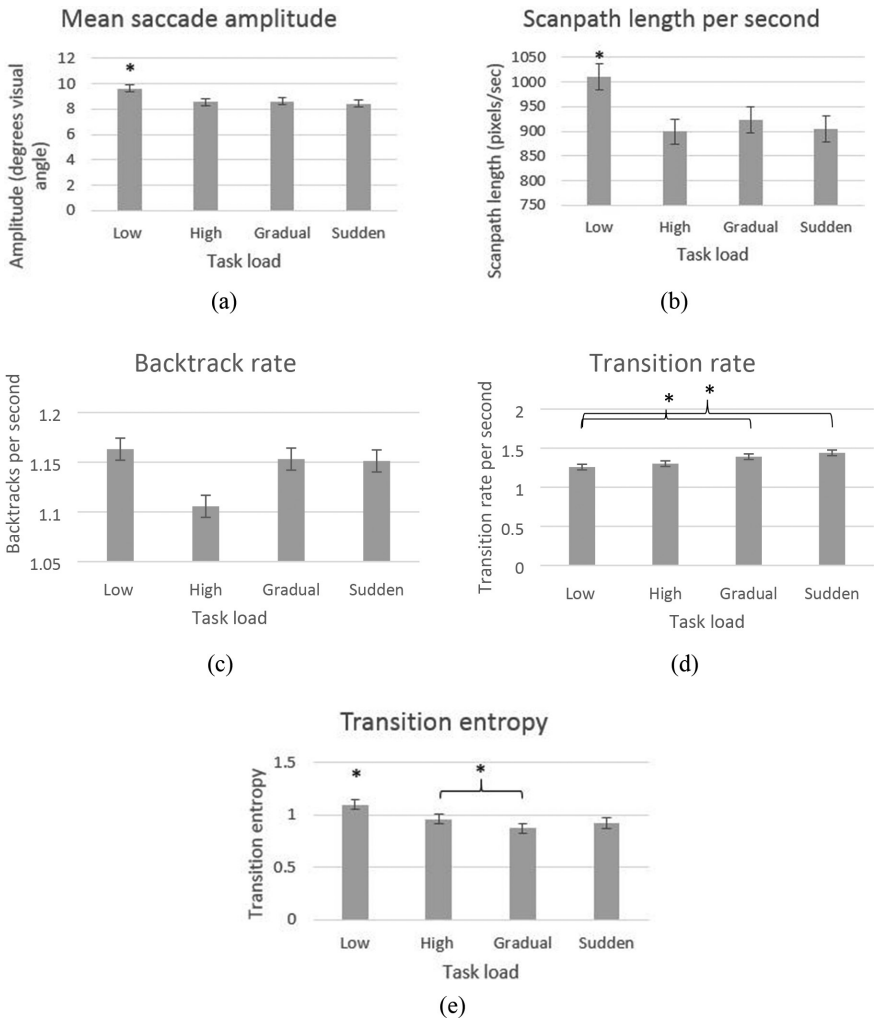


Figure 8. Directness metrics results for each task load condition: (a) scanpath length, (b) mean saccade amplitude, (c) backtrack rate, (d) rate of transitions, and (e) transition entropy.

by eye movements, are affected by gradual and sudden workload transitions when compared to constant low or high workload in a multitasking environment. The performance results obtained in this experiment suggest that there was no significant difference between performance in the gradual and sudden workload conditions. Contrary to our expectations, this suggests that sudden workload transitions are not more detrimental to performance than gradual workload transitions.

The analysis of the primary and secondary tasks provided further insights into the effects of workload transitions. For the primary task, high workload led to the worst primary task performance, which is consistent with the literature on the dangers of high workload (Cain, 2007). The workload transition conditions—both gradual and sudden—led to primary task performance that was in between low and high workload. However, the performance results observed for the secondary task differ from those of the primary task. For the secondary task, low and high workload resulted in the worst performance, whereas the transition conditions resulted in the

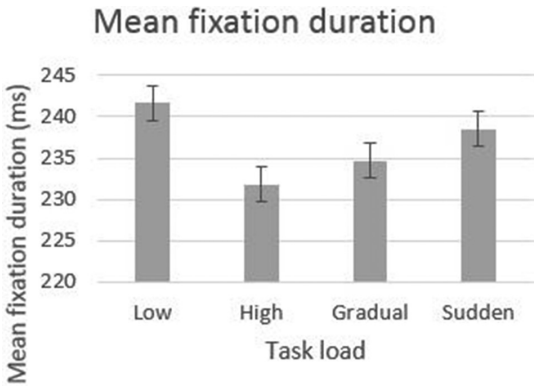


Figure 9. Duration metric results for each task load condition.

best performance. In other words, while task load was fluctuating between low and high, whether suddenly or gradually, secondary task response accuracy was higher compared to when workload was held constant at high or low. In general, the results were contrary to our expectation that, for both the primary and secondary tasks, performance would be increasingly worse in the low, gradual, sudden, and high workload scenarios. Nor do the results support our expectation that there would be a detrimental effect of sudden workload transitions, as found by researchers in other contexts (Cox-Fuenzalida, 2007; Kochan et al., 2004). However, the results are consistent with resource-allocation-based theories of attention, such as Wickens'

TABLE 2: Summary of the Eye Tracking Results

Metric	Result	Summary
Spread Metrics		
Convex hull area (pixels^2)	Not significant	H1: As performance deteriorates, these metrics will exhibit increased spread Findings: Spread of fixations is highest in low task load and lowest in gradual and sudden task load
Spatial density	• Highest value in low task load compared to gradual and sudden task load	
Stationary entropy	• Highest value in low task load • Higher value in high task load than gradual task load	
Directness Metrics		
Mean saccade amplitude (pixels)	• Highest value in low task load	H2: As performance deteriorates, these metrics will exhibit less directness and efficiency Findings: Low task load is the least efficient condition overall, but it has fewer transitions than the sudden and gradual task load conditions
Scanpath length per second (pixels)	• Highest value in low task load	
Backtrack rate (s^{-1})	Not significant	
Transition rate (s^{-1})	• Lowest value in low task load compared to gradual and sudden task load	
Transition entropy	• Highest value in low task load • Higher value in high task load than gradual task load	
Duration Metric		
Mean fixation duration (ms)	Not significant	H3: As performance deteriorates, there will be increased mean fixation duration Findings: Mean fixation duration is not affected by workload or workload transitions

(2002) multiple resource theory and Young and Stanton's (2002) malleable attentional resources theory. These theories posit that very low and very high workload can lead to performance decrements, whereas there exists an optimal middle-ground level of task load that elicits the best performance. Nevertheless, the results are surprising given that participants were asked to prioritize the primary task over the secondary one. The results confirm that workload transitions can affect how people perform their tasks in a way that is different from just high or low workload (Cox-Fuenzalida, 2007; Cumming & Croft, 1973; Goldberg & Stewart, 1980), even in a more complex, multitasking environment. More specifically, the results suggest that the transition conditions of sudden and gradual workload resulted in the participants adjusting their priorities and strategies as part of a multitasking operation.

The discrepancies with the previous literature serve to underline the importance of our main aim in this study, which is to use eye tracking metrics to better understand how these performance effects come about. In general, results were consistent with hypotheses H1 and H2 that there would be increased spread and less directness, respectively, with worse task performance. It is interesting, though, that the task performance that was best reflected in the eye tracking metrics was the secondary task, which relates mainly to transitioning between AOIs. It would appear that the transition between AOIs or the sequence of AOIs is an aspect that should be regularly explored in such studies, contrary to what is currently the case.

However, the eye tracking metrics reflected the poor performance mainly in the low but not high workload condition, although performance was equally poor in both. Under low task load, the *spread* metrics suggest that, in terms of where participants are looking, it appeared that participants were covering wider and more varied areas of the display. The *directness* metrics suggest that, in terms of efficiency, participants were scanning less efficiently under low task load and were more focused and efficient under the other task load conditions, especially in the gradual and sudden conditions. Only with stationary and transition entropy was the high workload condition also significantly different than the gradual workload condition, suggesting that these metrics are particularly sensitive to changes between constant and transitioning workload and would be recommended for studies in this context. This could be because these metrics are based on Markov chains, a stochastic process in which the next state (in this case, the next AOI that is fixated) depends only on the current state (current AOI being fixated). By modeling the scanpath in this way, it appears that one can very well capture differences in transition patterns. Moreover, the only exception to the increased directness under worse performance was observed in terms of the transition rate: sudden and gradual workload transitions exhibited higher transition rates compared to when workload was held constant at low. Finally, the only hypothesis that our findings fully contradicted was H3, which is related to the duration metric; results suggest that there was no effect of workload transitions on how long participants were looking at certain areas. This is surprising given how frequently mean fixation duration is used in the context of workload (e.g., Schulz et al., 2011). It could be that there is no issue of discriminating information in this study, although this would need to be further explored.

In summary, these metrics seem to suggest that low workload results in more spread in terms of where people are scanning, but eye movements become on average less efficient. On the other hand, in the workload transition conditions and under high workload, there is less spread, but higher efficiency. This can be explained by the fact that, under low task loads, participants had more time to scan the entire display without having to be systematic in terms of where they were looking. Although the greater spread may have benefitted the primary task performance, the lack of a systematic approach to scanning adversely affected secondary task performance. This nuance only serves to emphasize the importance of incorporating directness metrics to thoroughly understand attention allocation.

The results can be further interpreted in terms of the two-dimensional framework for the role of attention proposed by Trick et al.

(2004). In the proposed framework, there are two dichotomies: controlled versus automatic and endogenous versus exogenous. In relation to this study, participants in the low workload condition would be operating in the controlled-exogenous mode of attention. This mode promotes more open-ended exploration of the environment/display that is supported by spread metrics observed here under low task load which is the default. However, Trick et al. (2004) note that even though attention in this mode is conscious and voluntary, it can interfere with secondary task performance, which was found to be the case here. Using the same framework, the increased task demands associated with the workload transitions and high workload conditions most closely align with the controlled-endogenous mode, where behavior is deliberate and goal-driven. This is supported by the fact that the increased task demands lead to a need to balance priorities under the dual-task paradigm, which may have come at a cost to primary task performance.

Given that the tasks in this study involved a visual search task (i.e., target detection task), the findings can also be interpreted in light of the considerable visual search literature (see Wolfe and Horowitz (2017) for an overview). Models of visual search generally characterize search as a multistage process, with an early parallel stage obtaining basic information about the wider display area and a subsequent serial, slower stage getting more detailed information, but from limited display areas (Wolfe, 1994). This ability to obtain basic information from the periphery is vital to the visual search process (Rosenholtz, 2016; Wolfe et al., 2017). However, the presence of crowding or clutter, as is the case in this study as the number of active UAVs increases, makes it more difficult to recognize and discriminate objects in the periphery. This made it necessary for participants to saccade to and foveate items detected in the periphery to be able to identify their properties (Wolfe & Whitney, 2014). In the presence of less crowding in low workload, items in the periphery may have been more salient and thus drew participants' visual attention to wide areas of the screen. This saccade to wide areas may have been what led to larger spread in low workload, with the resulting fixations on the targets helping to improve performance in the primary task, which required detailed discrimination. This theory is also consistent with the higher mean saccade amplitude in low workload, which suggests that items in the wider periphery were detected and a (large) saccade was initiated to that item. This increased spread may not have been necessary or helpful for the secondary tasks, though, where peripheral vision may have been enough to understand what needs to be attended to.

Moreover, it has been shown that people can plan saccades in advance of search, although, if they have enough time and an item is salient enough, their planned scanpath can be altered (De Vries et al., 2014). In lower workload, it appears that there was more time for participants to change their planned scanpath when a salient object appeared in the periphery. This would explain the more random and less efficient sequence of eye movements in low workload, with participants' eye movements drawn to different areas at random. In the other workload conditions (high, sudden, and gradual), all of which involved higher workload, it would appear that participants tended to stick to a selected sequence and not deviate from that. This more systematic approach would also explain the increased transition rate in these three conditions, with participants seemingly deciding to move from one area of the screen to another in a planned fashion. The findings here demonstrate the importance of considering multiple types of eye tracking metrics to understand the effects of workload on visual attention allocation as thoroughly as possible.

CONCLUSION

There are three main takeaways from these findings. The first is that, contrary to expectations, sudden workload transitions do not seem to be more detrimental than gradual workload transitions. There does not seem to be any evidence of the "surprise" of a sudden transition leading to any decrements in performance, or any changes in attention allocation, either. The element of surprise and unexpected events may have detrimental effects in certain contexts such as aviation (Wickens, 2001); however, as part of this study, the element of surprise that may

have occurred with a sudden workload shift may have been diluted as there was more than one workload transition over the course of the 15 min scenario. If this study were replicated with just one sudden increase, the element of surprise present with the sudden task load condition could be accentuated and lead to decrements in performance.

At the same time, the second conclusion that can be drawn from these findings is that transitioning workload could affect how people prioritize tasks in a multitasking environment. It would seem that switching task load between low and high workload leads to participants changing their strategy. Transitioning workload cannot be treated simply as a "middle" stage between low and high workload, with the expectation that performance will be averaged. Instead, the findings here should be taken into account in the design of tasks within safety-critical, multitasking domains.

The third and final conclusion is the importance of using a combination of eye tracking metrics, together with a knowledge of the context, to understand attention allocation. In particular, the spread and directness metrics, which are much less used as compared to duration measures (e.g., Coyne et al., 2017; Foy & Chapman, 2018), are the ones that provided the more insightful results. Specifically, the combination of spread metrics (namely, spatial density and stationary entropy) and directness metrics (namely, mean saccade amplitude, scanpath length per second, transition rate, and transition entropy) was sensitive to differences and changes in task demands. The directness metrics in general are the ones that helped explain how participants did worse on the secondary task in low workload, contrary to all expectations. Stationary and transition entropies were particularly insightful, suggesting that they should be used more in studies on transitioning workload. Given that stationary entropy reflects just the switching between different areas of the screen, regardless of what the exact task is within each AOI, these metrics could still provide insight into participants' task switching behavior when other tasks are used.

In terms of more long-term future applications, these metrics can potentially be used as the basis for an adaptive display, where information is updated to suit users' needs in real time to support SA (e.g., Monfort et al., 2016). To this end, it is important that the appropriate metrics are used to account for different workload conditions. Once the system detects that the user is struggling with high workload or workload transitions, display adjustments can then be triggered in near real time before significant performance breakdowns occur. This could be extremely valuable in safety- and time-critical domains, such as military operations, aviation, or process control. Using eye tracking allows for the triggering of display adjustments as early as a few seconds into a task (Moacdieh & Sarter, 2017). At the same time, this approach would mean that there would be no need to trigger any unnecessary adjustments that might confuse or distract the user. This approach of *adjusting only when needed* is a cornerstone of adaptive and intelligent displays. However, note that it is critical that these metrics are not considered in isolation and should also take context into account. What works for one task may not work for all tasks in a complex, data-rich domain where operators are tasked with various responsibilities.

Although the results of this empirical work are informative, they are not free of limitations. First, generalizing these results is limited to domains with similar task structure, or ones that consist of monitoring several distinct areas of a screen and where increasing the number of items to attend to is directly proportional to increasing task load. The assumption is also that all the tasks are visual, not auditory. In its simplest form, this experiment involves search and monitoring multiple areas of the screen, and the fact that the eye tracking metrics were able to reflect this transitioning behavior well—namely, with the entropy metrics—suggests that any kind of transitioning between AOIs could also be captured, even in a different context. However, more research is needed to improve the external validity of the study and establish which eye tracking metrics best reflect fluctuations in workload in other multitasking scenarios.

It would also be interesting to look at shorter scenarios in addition to longer time periods, as was done here. Analyzing a single time transition at a time could also provide insight into how operators' attention is affected immediately after transitions. Furthermore, the sampling rate of the

eye tracker used in this study (60 Hz) is not ideal for the study of saccade amplitude, although it makes no difference to the detection of fixations (Leube et al., 2017). To further explore the use of that metric, an eye tracker with a higher sampling rate (e.g., 120 Hz) will have to be used in future studies. Also, it would be interesting to further explore how results would change if participants were completely unaware of any changes in workload. In that case, the focus would fully be on the element of surprise.

AUTHOR CONTRIBUTIONS

N. Moacdieh worked on the design of the experiment, the write-up, and the analysis of results. S. Devlin contributed to the design of the experiment, running the experiment, and the write-up. H. Jundi worked on the eye tracking code, the analysis of the metrics, and the write-up of results. S. Riggs worked on the design of the experiment, the write-up, and the analysis of results

ACKNOWLEDGMENTS

This study was supported in part by the National Science Foundation (NSF grant: #1566346; Program Manager: Dr. Ephraim Glinert). The authors would like to thank Aakash Bhagat for his help in developing the simulator for this study.

ORCID iDs

Nadine Marie Moacdieh https://orcid.org/0000-0002-7677-7946

Sara Lu Riggs https://orcid.org/0000-0002-0112-9469

REFERENCES

Bergen, J. R., & Julesz, B. (1983). Parallel versus serial processing in rapid pattern discrimination. *Nature*, *303*(5919), 696–698. https://doi.org/10.1038/303696a0

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., Olmstead, R. E., Tremoulet, P. D., & Craven, P. L. (2007). EEG correlates of task engagement and mental task load in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine*, *78*(5 Suppl), B231–B244.

Bliss, J. P., & Dunn, M. C. (2000). Behavioural implications of alarm mistrust as a function of task workload. *Ergonomics*, *43*(9), 1283–1300. https://doi.org/10.1080/001401300421743

Boyer, M., Cummings, M. L., Spence, L. B., & Solovey, E. T. (2015). Investigating mental workload changes in a long duration supervisory control task. *Interacting with Computers*, *27*(5), 512–520. https://doi.org/10.1093/iwc/iwv012

Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology*, *42*(3), 361–377. https://doi.org/10.1016/0301-0511(95)05167-8

Cain, B. (2007). *A review of the mental workload literature. Technical report*, Defence Research and Development Toronto.

Coral, M. P. (2016). A*nalyzing cognitive workload through eye-related measurements: A meta-analysis* [Doctoral dissertation]. Retrieved from Wright State University database.

Cox-Fuenzalida, L.-E. (2007). Effect of workload history on task performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(2), 277–291. https://doi.org/10.1518/001872007X312496

Cox-Fuenzalida, L.-E., Beeler, C., & Sohl, L. (2006). Workload history effects: A comparison of sudden increases and decreases on performance. *Current Psychology*, *25*(1), 8–14. https://doi.org/10.1007/s12144-006-1012-6

Coyne, JT., Sibley, C., Sherwood, S., Foroughi, CK., Olson, T., & Vorm, E. (2017, July). *Assessing workload with low cost eye tracking during a supervisory control task* [Conference session]. International Conference on Augmented Cognition, Vancouver, Canada, Springer, 139–147.

Cumming, R. W., & Croft, P. G. (1973). Human information processing under varying task demand. *Ergonomics*, *16*(5), 581–586. https://doi.org/10.1080/00140137308924548

De Rivecourt, M., Kuperus, M. N., Post, W. J., & Mulder, L. J. M. (2008). Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. *Ergonomics*, *51*(9), 1295–1319. https://doi.org/10.1080/00140130802120267

De Vries, J. P., Hooge, I. T. C., & Verstraten, F. A. J. (2014). Saccades toward the target are planned as sequences rather than as single steps. *Psychological Science*, *25*(1), 215–223. https://doi.org/10.1177/0956797613497020

Di Nocera, F., Camilli, M., & Terenzi, M. (2007). A random glance at the flight deck: Pilots' scanning strategies and the real-time assessment of mental workload. *Journal of Cognitive Engineering and Decision Making*, *1*(3), 271–285. https://doi.org/10.1518/155534307X255627

Di Stasi, L. L., Antolí, A., & Cañas, J. J. (2013). Evaluating mental workload while interacting with computer-generated artificial environments. *Entertainment Computing*, *4*(1), 63–69. https://doi.org/10.1016/j.entcom.2011.03.005

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(3), 474–486. https://doi.org/10.1518/001872006778606822

Dixon, S. R., Wickens, C. D., & Chang, D. (2005). Mission control of multiple unmanned aerial vehicles: A workload analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *47*(3), 479–487. https://doi.org/10.1518/001872005774860005

Duchowski, A. T. (2007). *Eye tracking methodology: Theory and practice*. Springer.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 32–64. https://doi.org/10.1518/001872095779049543

Endsley, MR. (1988). *Design and evaluation for situation awareness enhancement* [Paper presentation]. Proceedings of the Human Factors Society Annual Meeting, Los Angeles, CA, 97–101.

Eyetracking. (2011). Hardware: Eye tracking systems. http://www.eyetracking.com/Hardware/Eye-Tracker-List

Feitshans, G., Rowe, A., Davis, J., Holland, M., & Berger, L. (2008). *Vigilant spirit control station (VSCS): The face of COUNTER* [Conference session]. AIAA Guidance, Navigation and Control Conference and Exhibit, Honolulu, HI, American Institute of Aeronautics and Astronautics. 6309

Foy, H. J., & Chapman, P. (2018). Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation. *Applied Ergonomics*, *73*, 90–99. https://doi.org/10.1016/j.apergo.2018.06.006

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: Methods and constructs. *International Journal of Industrial Ergonomics*, *24*(6), 631–645. https://doi.org/10.1016/S0169-8141(98)00068-7

Goldberg, R. A., & Stewart, M. R. (1980). Memory overload or expectancy effect? 'Hysteresis' revisited. *Ergonomics*, *23*(12), 1173–1178. https://doi.org/10.1080/00140138008924824

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95–112. https://doi.org/10.1007/BF02289823

Hampson, R. E., Opris, I., & Deadwyler, S. A. (2010). Neural correlates of fast pupil dilation in nonhuman primates: Relation to behavioral performance and cognitive workload. *Behavioural Brain Research*, *212*(1), 1–11. https://doi.org/10.1016/j.bbr.2010.03.011

Hancock, P. A., Williams, G., & Manning, C. M. (1995). Influence of task demand characteristics on workload and performance. *The International Journal of Aviation Psychology*, *5*(1), 63–86. https://doi.org/10.1207/s15327108ijap0501_5

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139–183.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Hoover, A., Singh, A., Fishel-Brown, S., & Muth, E. (2012). Real-time detection of workload changes using heart rate variability. *Biomedical Signal Processing and Control*, *7*(4), 333–341. https://doi.org/10.1016/j.bspc.2011.07.004

Huey, B. M., & Wickens, C. D. (1993). *Workload transition: Implications for individual and team performance*. National Academy Press.

Hwang, S. L., Yau, Y. J., Lin, Y. T., Chen, J. H., Huang, T. H., Yenn, T. C., & Hsu, C. C. (2008). Predicting work performance in nuclear power plants. *Safety Science*, *46*(7), 1115–1124. https://doi.org/10.1016/j.ssci.2007.06.005

Kochan, JA., Breiter, EG., & Jentsch, F. (2004, September). *Surprise and unexpectedness in flying: Database reviews and analyses* [Conference session]. Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 48, No. 3, pp. 335–339), Los Angeles, CA.

Krejtz, K., Szmidt, T., Duchowski, A. T., & Krejtz, I. (2014, March). *Entropy-based statistical analysis of eye movement transitions* [Symposium]. Proceedings of the Symposium on Eye Tracking Research and Applications (pp. 159–166), Safety Harbor, FL, ACM.

Leube, A., Rifai, K., & Wahl, S. (2017). Sampling rate influences saccade detection in mobile eye tracking of a reading task. *Journal Eye Movement Ressearch*, *10*(3), 4–14.

Matthews, M. L. (1986). The influence of visual workload history on visual performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *28*(6), 623–632. https://doi.org/10.1177/001872088602800601

Matthews, G., Reinerman-Jones, L. E., Barber, D. J., & Abich IV, J. (2015). The psychometrics of mental workload: Multiple measures are sensitive but divergent. *Human Factors*, *57*(1), 125–143.

Moacdieh, N., & Sarter, N. (2015a). Clutter in electronic medical records: Examining its performance and attentional costs using eye tracking. *Human Factors*, *57*(4), 591–606.

Moacdieh, N., & Sarter, N. (2015b). Display clutter: A review of definitions and measurement techniques. *Human Factors*, *57*(1), 61–100.

Moacdieh, N. M., & Sarter, N. (2017). Using eye tracking to detect the effects of clutter on visual search in real time. *IEEE Transactions on Human-Machine Systems*, *47*(6), 896–902. https://doi.org/10.1109/THMS.2017.2706666

Monfort, SS., Sibley, CM., & Coyne, JT. (2016, May). *Using machine learning and real-time workload assessment in a high-fidelity UAV simulation environment* [Conference session]. In Next-Generation Analyst IV (Vol. 9851, p. 98510B), Baltimore, MD, International Society for Optics and Photonics, San Diego, CA.

Morgan, J. F., & Hancock, P. A. (2011). The effect of prior task loading on mental workload: An example of hysteresis in driving. *Human Factors*, *53*(1), 75–86.

National Transportation Safety Board. (1978). Aircraft Accident Report - United Airlines, Inc., McDonnell-Douglas DC-8-61, N8082U. Portland, Oregon.

Poole, A., & Ball, L. J. (2005). Eye tracking in human-computer interaction and usability research. In C. Ghaoui (Ed.), *Encyclopedia of human computer interaction (pp. 211–219)*. Idea Group.

Prytz, E. G., & Scerbo, M. W. (2015). Changes in stress and subjective task load over time following a task load transition. *Theoretical Issues in Ergonomics Science*, *16*(6), 586–605. https://doi.org/10.1080/1463922X.2015.1084397

Rantanen, E. M., & Goldberg, J. H. (1999). The effect of mental task load on the visual field size and shape. *Ergonomics*, *42*(6), 816–834. https://doi.org/10.1080/001401399185315

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In P. A. Hancock & N. Meshkati (Eds.), *Advances in psychology* (Vol. 52, pp. 185–218). North-Holland.

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual Review of Vision Science*, *2*(1), 437–457. https://doi.org/10.1146/annurev-vision-082114-035733

Rouse, W. B., Edwards, S. L., & Hammer, J. M. (1993). Modeling the dynamics of mental workload and human performance in complex systems. *IEEE Transactions on Systems, Man, and Cybernetics*, *23*(6), 1662–1671. https://doi.org/10.1109/21.257761

Rozado, D., & Dunser, A. (2015). Combining EEG with pupillometry to improve cognitive workload detection. *Computer*, *48*(10), 18–25. https://doi.org/10.1109/MC.2015.314

Savage, S. W., Potter, D. D., & Tatler, B. W. (2013). Does preoccupation impair hazard perception? A simultaneous EEG and eye tracking study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *17*, 52–62. https://doi.org/10.1016/j.trf.2012.10.002

Schulz, C. M., Schneider, E., Fritz, L., Vockeroth, J., Hapfelmeier, A., Wasmaier, M., Kochs, E. F., & Schneider, G. (2011). Eye tracking for assessment of workload: A pilot study in an anaesthesia simulator environment. *British Journal of Anaesthesia*, *106*(1), 44–50. https://doi.org/10.1093/bja/aeq307

Solovey, E. T., Zec, M., Garcia Perez, E. A., Reimer, B., & Mehler, B. (2014, April). *Classifying driver workload using physiological and driving performance data: Two field studies* [Conference session]. Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (pp. 4057–4066), Toronto, Canada, ACM.

Taylor, R. (1990). *Situational Awareness Rating Technique (SART): The development of a tool for aircrew systems design*[Symposium]. AGARD, Situational Awareness in Aerospace Operations 17 P(SEE N 90-28972 23-53), Neuilly-Sur-Seine, France.

Teichner, W. H. (1974). The detection of a simple visual signal as a function of time of watch. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *16*, 339–352. https://doi.org/10.1177/001872087401600402

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Trick, L. M., Enns, J. T., Mills, J., & Vavrik, J. (2004). Paying attention behind the wheel: A framework for studying the role of attention in driving. *Theoretical Issues in Ergonomics Science*, *5*(5), 385–424. https://doi.org/10.1080/14639220412331298938

Van Benthem, K. D., Herdman, C. M., Tolton, R. G., & LeFevre, J. A. (2015). Prospective memory failures in aviation: Effects of cue salience, workload, and individual differences. *Aerospace Medicine and Human Performance*, *86*(4), 366–373. https://doi.org/10.3357/AMHP.3428.2015

Veltman, J. A., & Gaillard, A. W. K. (1996). Physiological indices of workload in a simulated flight task. *Biological Psychology*, *42*(3), 323–342. https://doi.org/10.1016/0301-0511(95)05165-1

Wickens, C. D. (1992). Workload and situation awareness: An analogy of history and implications. *Insight: The Visual Performance Technical Group Newsletter*, *14*(4), 1–3.

Wickens, C. D. (2001, May). *Attention to safety and the psychology of surprise* [Symposium]. Proceedings of the 2001 Symposium on Aviation Psychology, Columbus, OH, The Ohio State University

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177. https://doi.org/10.1080/14639220210123806

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238. https://doi.org/10.3758/BF03200774

Wolfe, B., Dobres, J., Rosenholtz, R., & Reimer, B. (2017). More than the useful field: Considering peripheral vision in driving. *Applied Ergonomics*, *65*, 316–325. https://doi.org/10.1016/j.apergo.2017.07.009

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Publishing Group*, *1*, 1–8. https://doi.org/10.1038/s41562-017-0058

Wolfe, B. A., & Whitney, D. (2014). Facilitating recognition of crowded faces with presaccadic attention. *Frontiers in Human Neuroscience*, *8*, 103. https://doi.org/10.3389/fnhum.2014.00103

Young, M. S., & Stanton, N. A. (2002). Malleable attentional resources theory: A new explanation for the effects of mental underload on performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *44*(3), 365–375. https://doi.org/10.1518/0018720024497709

Nadine Marie Moacdieh is an assistant professor in Industrial Engineering and Management at the American University of Beirut. She obtained her PhD in Industrial and Operations Engineering from the University of Michigan, Ann Arbor in 2015.

Shannon P. Devlin is a PhD candidate in Systems Engineering in the Department of Engineering Systems and Environment at the University of Virginia. She received her MS in Industrial Engineering from Clemson University in 2018.

Hussein Jundi is a Masters student in Data Engineering and Analytics at the Technical University of Munich. He received his BE in Industrial Engineering and Management from the American University of Beirut in 2018.

Sara Lu Riggs is an assistant professor in Systems Engineering in the Department of Engineering Systems and Environment at the University of Virginia. She obtained her PhD in Industrial and Operations Engineering from the University of Michigan, Ann Arbor in 2014.

Improving Traffic Incident Management Using Team Cognitive Work Analysis

Vanessa Cattermole-Terzic, Queensland Department of Transport and Main Roads, Australia; The University of Queensland, Australia, and **Tim Horberry**, Monash University, Australia; Coventry University, UK

Effective traffic incident management requires separate responder agencies, with different and sometimes competing priorities and purposes, to come together as a team. Their priorities include optimizing casualty outcomes, minimizing the disruption to the flow of traffic, and maintaining responder team safety. In this study, team Cognitive Work Analysis was used in a desktop exercise setting to analyze a complex traffic incident management exercise. The study investigated decisions made at the scene of an incident to determine system issues and system support solutions. Participants were all senior officers and decision makers in traffic incident management environments. Results indicated that team Cognitive Work Analysis was highly beneficial in determining gaps in team coordination, communication, and structures. Information regarding shared and not shared work elements between agencies highlighted novel coordination and education requirements within and between agencies, such as disparate priorities at the scene creating the risk of interoperability issues. Analyses of operational, coordination, and structural strategies offered new insights into the traffic incident management work domain and recommendations for improvements to the safety and performance of the overall traffic incident management system.

Keywords: cognitive work analysis, critical decision method, teamwork, traffic incident management

Address correspondence to Vanessa Cattermole-Terzic, Level 17, 61 Mary St, Brisbane, Australia, 4001, Vanessa.Z.Cattermole@tmr.qld.gov.au.

Journal of Cognitive Engineering and Decision Making
2020, Volume 14, Number 2, June 2020, pp. 152–173
DOI: 10.1177/1555343419882595

INTRODUCTION

Traffic incident management involves the coordinated response from emergency services, traffic agencies, and local government agencies to remove incidents and restore traffic capacity safely and efficiently (Charles, 2007b; Farradyne, 2000). Aside from improved road safety, effective traffic incident management reduces congestion costs, improves the reliability of all forms of transport, and reduces vehicle emissions. As an example, congestion costs in the United States in 2005 were estimated at 78.2 billion dollars, with 52% to 58% of congestion attributed to traffic incidents (Carson, 2010).

However, traffic incident management involves inherent risks and the eliminating these risks is difficult due to the complexity of traffic incident environments. The traffic incident management work environment is dynamic, and the characteristics at each incident vary. Examples of issues related to incident specifics include the resources available at the scene (including availability of equipment like lighting, variable message signs, vests, traffic cones and so on, but also including personnel—quantity and levels of expertise/specialization for requirements at the scene), communication at the scene (intra- and interagency), the physical characteristics of the scene (scene size, topography, weather conditions, time of day), and incident specifics (stability of vehicles, presence of hazardous materials [HAZMATs], number, and types of casualties; Charles, 2007).

Although the incident specifics can seem unique and random, experts report similarities across incidents and recognition of incident types (Klein, 1998). Therefore, despite the complexity of traffic incident work environments, one way to determine improvement opportunities for traffic incident management

is through an operator-centered focus. In a study by Cattermole, Horberry, Wallis, and Cloete (2014), emergency responders from police, fire, and traffic control agencies were asked to rate and describe their experiences of the greatest safety issues at traffic incident scenes. The highest rated issues related to interoperability between the emergency responder agencies. This finding is supported by Fiore and Salas (2004), who cited team coordination as a major issue for teams by researchers and a major aim of work in this area is to reduce the "process loss" of poor team coordination.

Teamwork involves the adaption of coordination strategies through closed-loop communication and a sense of collective orientation (Salas & Fiore, 2004). A good team in a temporally challenged, high stakes and dynamic environment such as in the traffic incident management environment requires a shared awareness of team goals, congruence between individual and team goals, and good coordination between team members conducting their separate tasks as part of the whole output (Charles, 2007). Training within agencies is strict, and professional development ongoing, so intra-agency teams display the characteristics described (Cattermole, Horberry, & Hassall, 2016). However, aside from occasional joint exercises, there is no training to better understand interagency roles and responsibilities at incident scenes. Also, due to the distributed and hectic nature of traffic incident management systems, team awareness is reduced, and team awareness is an important factor for successful collaboration (Gutwin & Greenberg, 2002).

The issue of interoperability at the scene of a traffic incident extends wider than the operational context. The traffic incident management environment can be thought of as a single system (Cattermole-Terzic, 2017). However, it is supported by policies and directives from separate agencies, departments, and industry, each developed with a focus on one aspect or group of the incident management system rather than the system as a whole. It is likely that some policies and practices will not be compatible. These incompatibilities may not have an immediate visible effect but may contribute to the potential for secondary incidents. Secondary incidents are incidents that occur after the initial incident and as a direct result of the changed conditions caused by the initial incident. The impact on secondary incidents includes loss of life, serious injury, and community, economic, and environmental costs. Secondary incidents also have a direct impact on emergency responder safety. For example, in the United States, an average of one police officer per month is killed in a roadside crash (Fischer, Krzmarzick, Menon, & Shankwitz, 2012).

The unique nature of each incident scene means that although training of standard operating procedures plays an important role in effecting optimal safety and output at traffic incident scenes, a large part of the work for teams at incidents requires problem solving, building knowledge, dynamic risk assessments, and flexibility. Agency training to build these skills at an intra- and interagency level may be one potential tool required to improve interoperability at incidents.

Frameworks from human factors have previously been used to map complex environments to improve safety and interoperability (e.g., Cooke, Gorman, Myers, & Duran, 2013; Flavell & Wellman, 1975; Kenny & La Voie, 1984; Klein & Wright, 2016; Langan-Fox, Code, & Langfield-Smith, 2000; McNeese, Rentsch, & Perusich, 2000; Rasmussen, 1983; Rentsch, Mello, & Delise, 2010; Scherer & Petrick, 2001; Shiffrin & Schneider, 1977). One such framework is Cognitive Work Analysis.

Cognitive Work Analysis

Cognitive Work Analysis is a framework that was originally developed by Jens Rasmussen (Rasmussen, 1983; Rasmussen, Pejtersen, & Schmidt, 1990). It is a theoretical framework that analyzes how people work in complex environments, with the aim of providing recommendations to improve system design (Vicente, 1999). Whereas other work analyses are descriptive (describing how work is currently done) or normative (describing how work should be done), Cognitive Work Analysis is a formative model (describing how work can be done; Naikar & Elix, 2016). It does this by identifying system controls and constraints over five phases of analysis. Each of the phases of

Cognitive Work Analysis focuses on different types of constraints. The initial phases focus on ecological elements and there is a gradual shift to more cognitive issues.

Ashoori and Burns (2013) modified the traditional five-phase approach of Cognitive Work Analysis into two sets of four. Work Domain Analysis, Control Task Analysis, Strategies Analysis, and Worker Competencies Analysis were paired with a parallel set of social or team models called team Work Domain Analysis, team Control Task Analysis, team Strategies Analysis, and team Worker Competencies Analysis. Using this modification, Ashoori and Burns successfully mapped teamwork and shared tasks, strategies to accomplish tasks, and the required qualifications of operators in effective medical teams. Therefore, aside from the ability of Cognitive Work Analysis to analyze complex systems where unanticipated events occur, the modification of the framework to focus on team interactions makes it ideal for analyzing traffic incident management. This study will adopt the Ashoori and Burns's (2013) adaptation, given its explicit focus on teamwork.

The four phases of team cognitive work analysis. A Work Domain Analysis provides a description of the constraints governing the functions and purpose of a particular work environment (Vicente, 1999). Team Work Domain Analysis investigates which team members have shared processes, components, and objectives within a work environment and also which elements only influence individuals. The goal of team Work Domain Analysis is to create a set of models that describe shared values, purpose, and priorities in the work environment.

Team Control Task Analysis investigates team activity and collaboration. One technique used in this process links the decision ladders of individual team members to determine team collaboration points in "decision wheels." Because this section of team Cognitive Work Analysis investigates how decisions are made, by whom, and when, it is particularly useful in determining team decision support requirements.

Team Strategies Analysis looks at how teams coordinate, form, and regroup to handle different tasks. Analysis of traffic incident management using team Strategies Analysis could yield valuable information for scenario testing. The work environment in traffic incident management is dynamic, but establishing lines of communication and determining team make up and procedures for different possibilities at incidents could improve synchronous collaboration and the overall effectiveness of traffic incident management.

Team Competency Analysis aims to determine a series of desirable competencies operators must possess to effectively work in a team. Team Competency Analysis extends the analysis of traditional Cognitive Work Analysis and includes a study of social competency, investigating the interpersonal skills required for effective teamwork. The analysis requires the identification of skills/rules/knowledge-based requirements in the work situations.

This study will use three of the four phases of team Cognitive Work Analysis—team Work Domain Analysis, team Control Task Analysis, and team Strategies Analysis—to investigate the collaboration requirements between responder agencies at traffic incidents. The framework was chosen primarily due to its focus on constraints and possibilities for behavior rather than describing how activities actually occur. The dynamic complexity of the traffic incident management environment would be impossible to encapsulate using descriptive or normative models.

In this Australian study, decision makers from Queensland Police Service, Queensland Fire and Emergency Services, and Royal Automotive Club Queensland's Traffic Response Unit participated in a complex desktop exercise. The results were analyzed using modified team Cognitive Work Analysis tools. The aim of this study was to establish collaboration points and to develop recommendations to improve intra- and interagency coordination, collaboration, and interoperability at traffic incident scenes.

METHOD

Participants

Participants for this study were required to be decision makers at traffic incidents and therefore needed to be senior officers. The researcher contacted the agencies to request for appropriate representation. The group Critical Decision Method (Klein, Calderwood, & Mac-

Figure 1. Map of the incident area provided to participants at the exercise.
Source. Nearmap.

gregor, 1989) for the desktop exercise included five participants—a Senior Traffic Response Officer, the Officer in Charge of the Forensic Crash Unit for the Queensland Police Service, two Queensland Fire and Emergency Services inspectors, and one Queensland Fire and Emergency Services Assistant Commissioner. All participants were male. The level of experience of the officers ranged from 15 to 32 years with an average experience of just over 25 years ($M = 25.8$ years, $SD = 5.9$) in the area of traffic incident management/emergency response.

Desktop Exercise and Photograph of the Incident Scene

The incident scenario was based on a major Australian incident on the Sydney F3 Sydney to Newcastle Freeway on April 12, 2010, when a 16-ton flatbed truck collided with the rear of a fully laden fuel carrier. The road was closed for a significant period, stranding motorists in some cases for more than 8 hr, resulting in significant media and political attention. For the exercise, the scenario was shifted to a South East Queensland location on the Pacific Motorway, just before the Logan River Bridge (Figure 1).

Procedure

The Critical Decision Method interview procedure was altered to suit a group environment to conduct a group desktop exercise. Participants were sent an incident scenario and a map of the incident location 2 days prior to the desktop exercise. They were asked to consider their agency's response to the given incident.

On the day, participants sat in a closed meeting room with a whiteboard at the University of Queensland. The desktop exercise was conducted using one interviewer. The interviewer had considerable experience working with the interviewing technique and also working in transport, police, and traffic incident environments. All participants gave informed consent to be part of the desktop exercise and to be audio-recorded throughout the process. The audio recording was taken to assist with later analyses. The desktop exercise process conformed to pre-approved University of Queensland ethics procedures. In total, the interview process for the desktop exercise took 3.5 hr.

Interview Process

A modification of the classic Critical Decision Method approach was used, applying four "sweeps" of the incident. In Critical Decision Method, sweeps are described as follows:

- Sweep 1: Incident familiarization: Participants were sent a scenario of an incident and asked to consider their agency response prior to attending the desktop exercise. At the exercise, the facilitator requested that the officers remain in the

mind-set of their own agency to prevent "group think" throughout the process.

At the beginning, the facilitator read through the incident details from beginning to end and then prompted each agency to begin with a description of what would happen from their agency's point of view, from beginning to end of the incident. Participants were encouraged to interject when they considered their agency would take over functions or collaborate. It was also the facilitator's role to keep the dialogue flowing and, as much as possible, in order of likely events/tasks.

Once the group finished, the interviewer retold the incident to the participants and wrote the details onto a whiteboard. The participants were encouraged to clarify and provide additional information where required so that a more complete description of the likely occurrences at the incident were represented on the whiteboard. At the end of this sweep, the interviewer and participants had a "shared view" of the incident. For this study, a shared view was considered to be reached when the interviewer retold the story according to what was written on the whiteboard and the participants agreed that all the details of likely occurrences of the incident were accurately described. This was considered to be the likely workflow for the scenario. At this point, the interviewer took photographs of the whiteboard to assist in the analysis stage of the study.

- Sweep 2: Timeline verification and decision point identification: The interviewer again retold the incident, encouraging the participants to organize the incident around a likely timeline. The interviewer then identified points where decisions would be made, and actions taken. These "decision points" formed the focus of the final two sweeps. Decisions were defined as points where the participant was required to act when more than one option was available to act upon. Decision points were identified by the interviewer. The interviewer clarified decision points with participants by confirming that the action was taken and then further understood through probe questions identifying that other options were available (e.g., questioning the participant about what a novice would have done in the situation, or what he would have done if some situational cues were altered).

It should be noted that all participants in the study considered the decision points as merely actions. Further probing and questions were needed at this point to determine information about how the participants determined what potential actions to take.

- Sweep 3: Deep probes: Probe questions were used to focus the participants on particular aspects of the cognitive processes and context behind the hypothetical decisions made by the experts at the incident. The questions included cue usage, prior knowledge, goals, expectations, and options.
- Sweep 4: Hypotheticals—What if . . .? In the final sweep, the participants were asked to shift their perspective to alternative views and outcomes. What if you were a novice in this situation? What if some particular aspect of the incident scene was different? This section examined the possibilities and consequences of other options and errors, and also extended the interviewer's understanding of the activation points for expert decisions.

Postinterview Activities

The desktop exercise was transcribed prior to analysis. To gain a richer understanding of the traffic incident management environment, the interviewer went on shift with participants from the Traffic Response Unit and Queensland Fire and Emergency Services. The data from the desktop exercise were then analyzed using three of the phases of team Cognitive Work Analysis and later validated by two senior emergency responders—one of the officers was a participant at the session, and one was external to the process.

RESULTS

Hypothesized Workflow for Agencies at the Incident

In Sweep 1 of the desktop exercise, participants identified the likely workflow for the scenario, depicted in Figure 2. In the workflow diagram, different roles and responsibilities are shown along the vertical axis, and the horizontal axis shows the work progress over time.

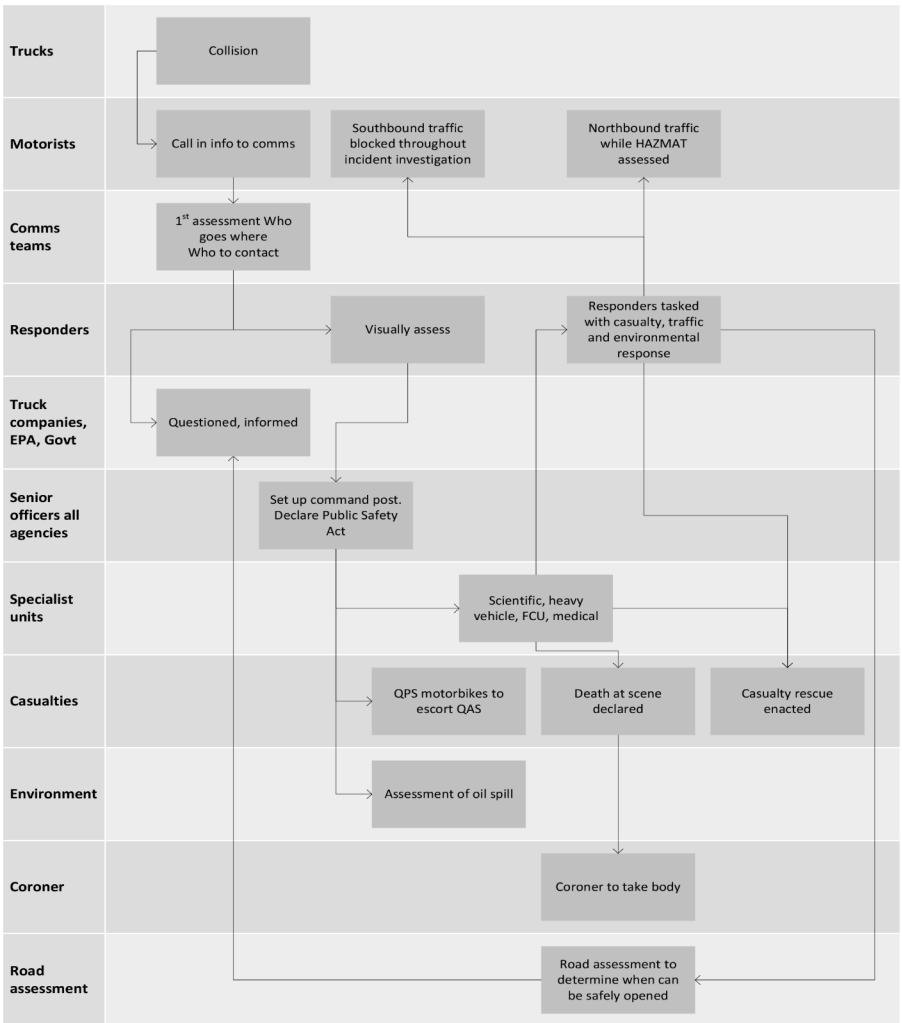


Figure 2. Workflow diagram of the hypothetical traffic incident.
Note. HAZMAT = hazardous materials; EPA = Environmental Protection Agency; FCU = Forensic Crash Unit; QPS = Queensland Police Service; QAS = Queensland Ambulance Service.

From this point, it is possible to analyze team structure, inter-team and intra-team interactions, shared work domain elements, and expertise using team Cognitive Work Analysis tools.

Team Work Domain Analysis

Team Work Domain Analysis investigates which team members have shared processes, components, and objectives within a work environment and also which elements only influence individuals. The first step in team Work Domain Analysis is to construct a regular Work Domain Model. A Work Domain Model is a diagram depicting the entire work domain, using five levels of abstraction for analysis. The top three levels consider the overall purpose of the work domain and what it is required to do. The bottom two levels consider capability and resource components. The five levels of abstraction are functional purpose, abstract function, generalized function, physical function, and physical form.

The different levels of the Work Domain Model are connected through a means-ends

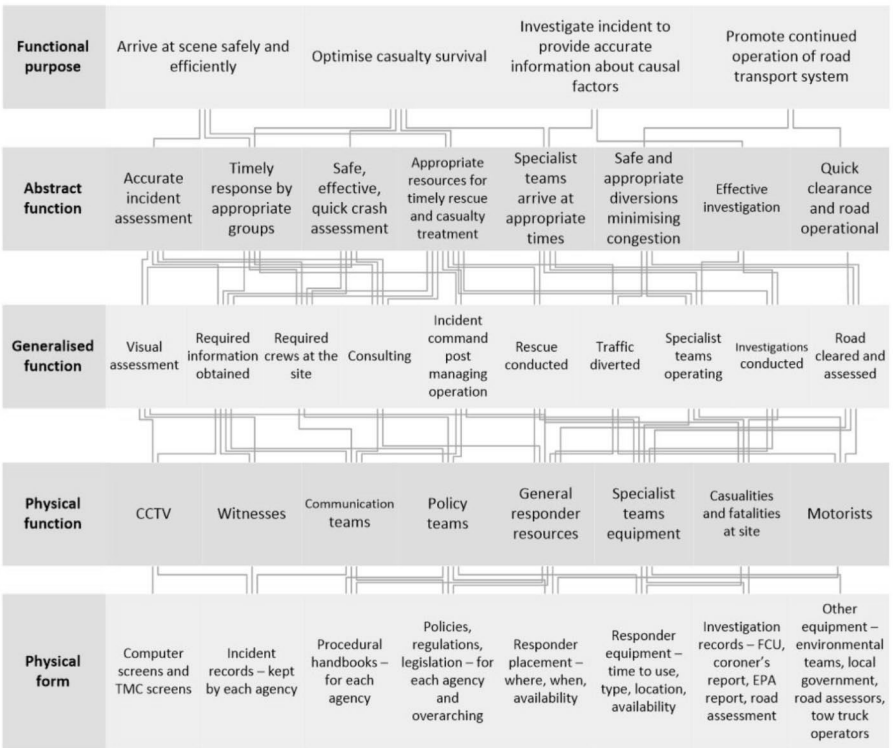


Figure 3. Work domain model for the incident.
Note. TMC = Traffic Management Center; FCU = Forensic Crash Unit; EPA = Environmental Protection Agency.

relationship. As an example, in the diagram below, at the physical form level, computer screens and traffic management center screens are required to visualize the CCTV footage (physical function level). The CCTV footage is required for the Brisbane Metropolitan Traffic Management Centre (BMTMC) to establish a visual assessment of the reported crash (generalized function). The visual assessment enables the BMTMC and traffic response unit to fulfill its priority of establishing an accurate incident assessment (abstract function), and the accurate assessment enables the traffic response to fulfill its higher level purpose for all personnel to arrive at the scene safely and efficiently. Figure 3 depicts the work domain model for the traffic incident scenario being analyzed.

At the functional purpose level, the abstraction hierarchy corresponds to work domain purposes. Four overall purposes were identified for traffic incident management at this incident at the functional purpose level: arrive at scene safely and efficiently, optimize casualty survival, investigate incident to provide accurate information about causal factors, and promote continued operation of the road transport system.

The abstract function level relates to the values, priorities, and principles of the teams at the incident. Eight processes were identified at this level: accurate incident assessment; timely response by appropriate groups; safe, effective, and quick accident assessment; appropriate resources for timely rescue; safe and appropriate diversions minimizing congestion; specialist teams arrive at appropriate times; effective investigations; quick clearance and road operational. The links between the abstract function level and the functional purpose level establish the why–how relation between the purposes of the work and the values and priorities.

Ten processes describe the generalized function level of the work domain model. These are visual assessment, required information obtained, required crews at the site, consulting,

incident command post managing operation, rescue conducted, traffic diverted, specialist teams operating, investigations conducted, and road cleared and assessed. This level identifies the main work processes at the incident and the links between generalized function level and the abstract function level identify the work domain processes that meet the values and priorities of the organizations at the scene.

The physical function level identifies the physical work domain resources. The eight resources identified were CCTV, witnesses, communication teams, policy teams, general responders, specialist teams, casualties, and motorists. Links between this level and the generalized function level identify the work domain resources.

At the physical form level, the physical characteristics of work domain resources are listed. The eight identified areas were computer screens and Traffic Management Center screens; incident records (kept by each agency); procedural handbook (for each agency); policies, regulations, and legislation (for each agency and overarching); responder placement (where, when, availability); responder equipment (time to use, type, location, availability); investigation records (Scenes of Crime Unit, Forensic Crash Unit, coroner's report, Environmental Protection Agency report, road assessment); and other equipment (for groups such as environmental teams, local government, road assessors, tow truck operators). The links between physical form and physical function levels identify the attributes of work domain resources.

The regular Work Domain Analysis identifies work domain purpose, values, work processes, and the physical elements of the work domain. However, it does not identify shared values or work processes. To better understand the responsibilities for each agency, a responsibility map was developed (Figure 4).

A significant level of shared responsibility at the intra- and interagency levels is evident at the incident.

To better identify shared/not shared elements across the work domain, the responsibility map was separated according to the four functional purposes identified for the incident. These were "arrive at the scene safely and efficiently," "optimize casualty survival," "investigate incident to provide accurate information about causal factors," and "promote continued operation of the transport system." As an example, the responsibility map for "optimize casualty survival" is described below.

Queensland Fire and Emergency Services and Queensland Ambulance Service are directly responsible for casualty outcomes at the scene (Figure 5). Differences in priorities for this purpose relate to the need for specialist teams from Queensland Fire and Emergency Services to be at the scene due to the nature of this particular incident. Conflicts can occur due to differing priorities at this incident, indicating a point at which Queensland Fire and Emergency Services and Queensland Ambulance Service need to establish shared understanding to prevent conflicts or counterproductive actions. The specialist team operation at the generalized function level is a clear area of Queensland Fire and Emergency Services activity that does not require coordination with Queensland Ambulance Service. All boundary objects at the physical form level are shared between agencies indicating a requirement to ensure they are compatible with the needs and purposes of both agencies.

To understand and analyze shared and individual purpose, values, processes, boundary objects, team structures and interactions in intra- and interagency context, the next step for team Work Domain Analysis is to build collaboration and abstraction tables.

Collaboration tables. The collaboration tables are divided into four levels—functional purpose, abstract function, generalized function, and physical function. As an example, this paper will outline the collaboration table for the functional purpose level. At this level, the Collaboration Table depicts roles and responsibilities that contribute to the work domain purpose. They are also useful to identify what collaboration should occur at scenarios versus the actuality of collaboration and coordination at incident scenes.

In Table 1, all communication teams and operational teams from the different agencies share the functional purpose of ensuring that the appropriate operational teams arrive at the scene

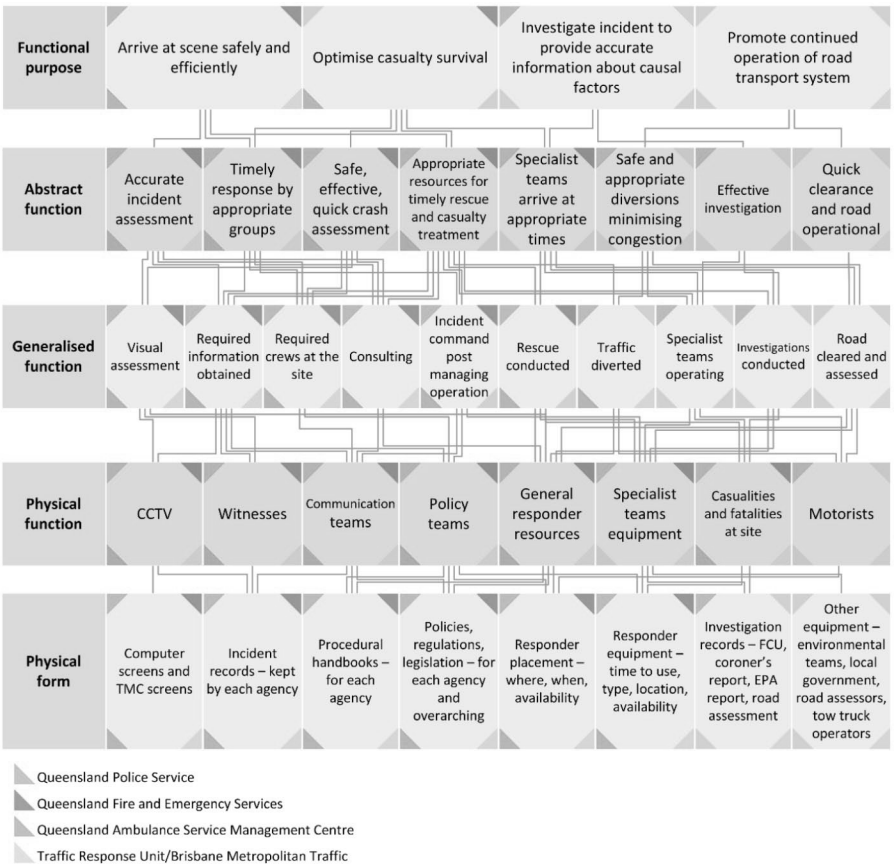


Figure 4. Responsibility map for the incident.
Note. FCU = Forensic Crash Unit; EPA = Environmental Protection Agency.

efficiently and safely. This suggests that coordination among the agencies should be apparent for activities that occur for this purpose. The other agency purposes at the scene are divided. Optimizing casualty outcomes is the shared purpose of Queensland Fire and Emergency Services and the Queensland Ambulance Service. Investigating the scene is the responsibility of the Queensland Police Service, the Environmental Protection Agency, and the Coroner. Ensuring that the road system is operational as quickly as possible is the responsibility of the Traffic Response Unit, tow truck operators, the Queensland Police Service, and the Department of Transport and Main Roads. This disparity of purpose in a safety critical environment sets the scene for possible interoperability issues. In the first row of Table 1, QPS refers to the Queensland Police Service, QFES is the Queensland Fire and Emergency Services, QAS is the Queensland Ambulance Service, BMTMC is the Brisbane Metropolitan Traffic Management Centre, TRU is the Traffic Response Unit, and DTMR is the Department of Transport and Main Roads.

Team Work Domain Analysis Findings

The purpose of this section of the analysis was to identify shared/not shared purposes, values, priorities, and principles. Ashoori, Burns, d'Entremont, and Momtahan (2014) found that team Work Domain Analysis provided information about shared elements of the information space in surgical teams. In support of the findings of Ashoori et al. (2014), this study also found that team Work Domain Analysis identified shared and not shared elements across

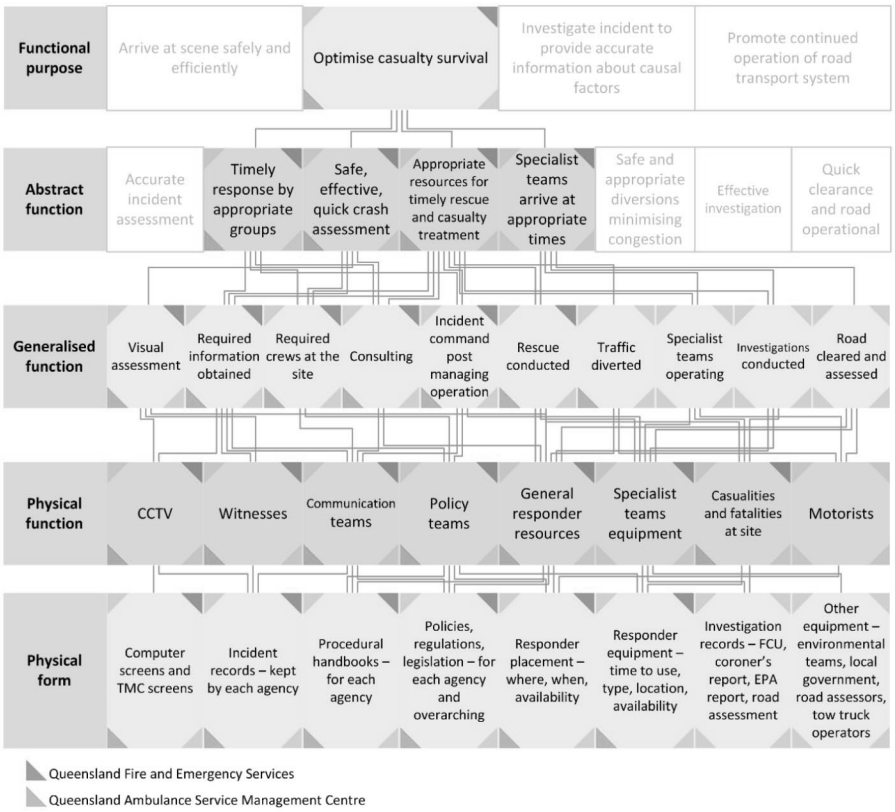


Figure 5. Responsibility map for "optimize casualty survival."
Note. TMC = traffic management center; FCU = Forensic Crash Unit; EPA = Environmental Protection Agency.

the emergency responder agencies, therefore identifying likely areas of weakness for interoperability.

More specifically, for the functional purpose of arriving at the scene safely and efficiently, all agencies have this purpose. At the abstract function level, however, whereas Queensland Fire and Emergency Services and the Queensland Ambulance Service prioritize optimal casualty outcomes at the scene, the Queensland Police Service prioritizes incident investigation and traffic management, and the Traffic Response Unit prioritizes traffic management, optimizing traffic flow through the area and road clearance and assessment. The different agency priorities in a high-stakes, complex and dynamic environment become a risk point for interoperability unless the responders from each agency have a shared awareness of the overall priorities of the Traffic Incident Management work domain, and the role each agency plays within it. However, in a study by Cattermole et al. (2014) investigating decision making and perspectives of emergency responders at traffic incidents, a shared awareness was not evident. For example, Queensland Fire and Emergency Services participants indicated that Queensland Police Service officers arrive too late to manage incident scenes, putting emergency responders from Queensland Ambulance Service and Queensland Fire and Emergency Services at greater risk and reducing the effectiveness of the incident response. Many did not seem aware that this is due to a different level of priority given to attending road crashes between the agencies.

Optimizing casualty outcomes at traffic incidents is a focus for Queensland Fire and Emergency Services and Queensland Ambulance

Service only. This indicates a need to ensure tight coordination between these agencies for this function, but little requirement for Queensland Police Service or Traffic Response Unit to understand the processes involved.

Incident investigation to determine causal factors at the incident is the purpose of Queensland Police Service and other specialty teams at the scene. Although it is not necessary for other teams to understand the investigation process, in the exercise, the impact of these teams taking several hours to attend the incident became evident. Although it is important for the site to remain untouched for criminal investigations, if chemicals are on the road for more than 3 hr, it is likely that the bitumen will need to be removed and therefore the road closed for several days. Although this aligns with the priorities of the investigation teams, it is in direct opposition to the priorities of the Traffic Response Unit and the Department of Transport and Main Roads.

Promoting continued operation of the road transport system is a functional purpose for the Queensland Police Service, Traffic Response Unit, Department of Transport, and tow truck operators. Aspects of this function require specialist activities for each agency individually, but joint functions should be coordinated so that policies, processes, and practices align and the agencies have a shared understanding of their joint roles in the function.

TABLE 1: Traffic Incident Management Collaboration Table at the Functional Purpose Level

Functional Purpose	Communication Team QPS	Communication Team QFES	Communication Team QAS	BMT-MC	QPS	QFES	QAS	TRU	Specialist Teams QFES	Specialist Teams QPS	Specialist Teams (Other)	Tow Trucks	DTMR/ Road Assessors
Arrive at scene safely and efficiently	X	X	X	X	X	X	X	X	X	X		X	
Optimize casualty survival						X	X		X				
Investigate incident to provide accurate information about causal factors					X					X			
Promote continued operation of the road system					X				X			X	X

Team Control Task Analysis

The team Control Task Analysis tools used to examine traffic incident management team structures, interactions, shared workflows, and boundary objects were decision wheels and the Contextual Activity Template for teams.

Decision wheels. Ashoori and Burns (2013) linked the decision ladders of individual team members to determine team collaboration points in "decision wheels." The maps become quite complex, so links are numbered for simplification. Using this technique, it is possible to create a decision wheel table and to determine whether the collaboration points were "synchronous" or "asynchronous." A synchronous collaboration point is one that is observed to occur efficiently

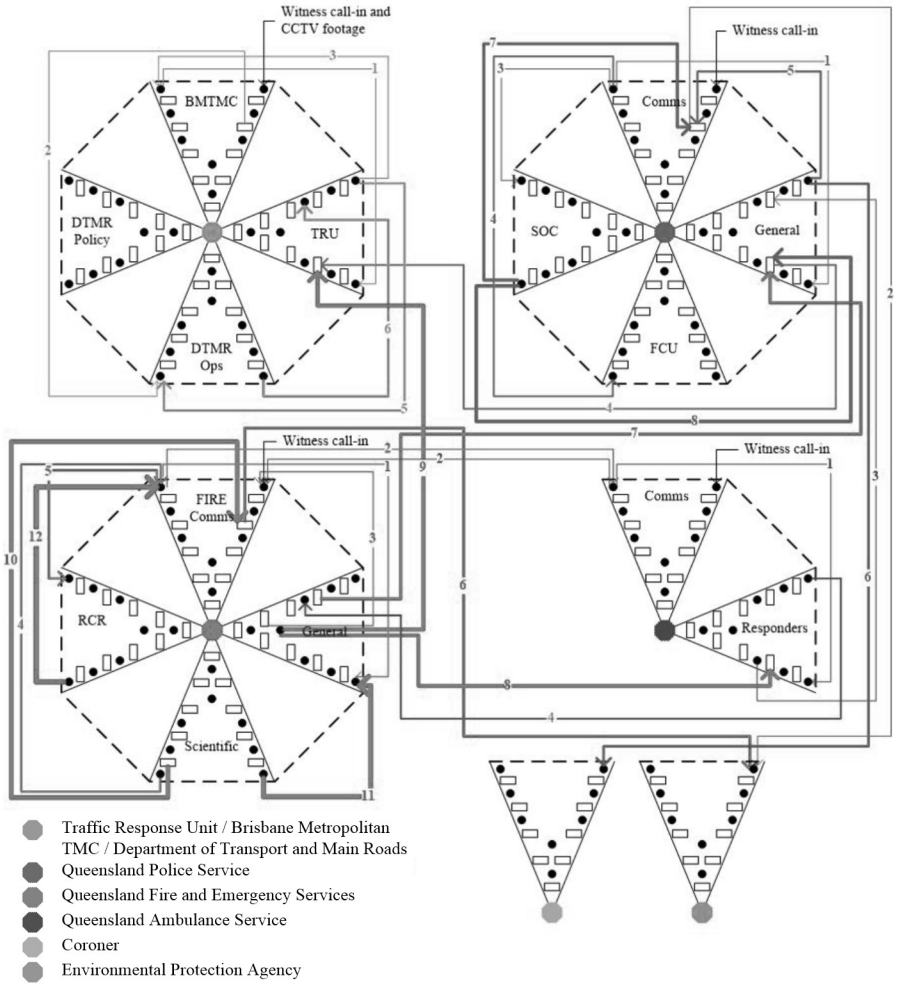


Figure 6. Decision wheel of traffic incident management incident.
Note. BMTMC = Brisbane Metropolitan Traffic Management Centre; DTMR = Department of Transport and Main Roads; TRU = Traffic Response Unit; SOC = Scenes of Crime Unit; FCU = Forensic Crash Unit; RCR = road crash rescue; TMC = traffic management center.

and effectively. An asynchronous collaboration is one that is observed to deviate from optimal effectiveness, efficiency, or safety. The decision wheels for the traffic incident management incident were complex and focused around Queensland Fire and Emergency Services due to the HAZMAT issues at the scene. Figure 6 depicts decision wheels for the incident. Each wheel represents an organization and each piece of "pie" within the wheel represents an actor/area/crew from the organization. The arrow pointing to a piece of pie represents the beginning point of a decision. For example, a line from witness call-ins goes to each of the agency communication teams, with the arrow pointing to the "activation" circle. Due to the hypothetical nature of the exercise, the flow through the decision ladder template cannot be depicted; however, the exit point gives information about when the communication teams make decisions about communicating with other units, and who they contact. The general complexity of traffic incident management is evident from the diagram also.

The numbers near each line identify the sequence of decisions and communications at

the intra- and interagency level. Colors of the arrows and numbers identify agencies—red for Queensland Fire and Emergency Services, blue for Queensland Police Service, purple for the Queensland Ambulance Service, and green for the Traffic Response Unit, BMTMC, and Department of Transport and Main Roads. Other groups represented in the diagram are the Environmental Protection Agency and tow truck operators.

The connections within and across decision wheels give information about communication and coordination requirements and the types of decisions and tasks required at each of the decision points. Due to the hypothetical nature of the scenario, the reported decision processes from the participants were primarily at the outer sections of the decision wheel. If this was actually the case, a focus on coordination would be less important. Decision processes that require knowledge-based decisions are more likely to require consideration of the impacts of other agencies and advice from agencies. As an example, in a previous study by Cattermole et al. (2016), a Queensland Fire and Emergency Services Officer required input from paramedics to determine the appropriate rescue operation. A series of changing circumstances communicated by paramedics altered the Station Officer's decisions about rescue operation requirements. It is highly likely that a complex incident like in this exercise would also contain several points where issues required the decision-making process to go to the higher levels of the decision ladder. Analysis of those points at the incident would identify crucial points for interagency communication and coordination.

Table 2 is an excerpt of a decision wheel table for the incident, using examples from Queensland Fire and Emergency Response to showcase the table's function. In the table, each of the decision points from each agency is listed by color-coded number. The table identifies the team making the decision, the tasks or information they were distributing, and the boundary object for the decision. In a real scenario, the table would also identify whether the decision was synchronous or asynchronous; however, in hypothetical situations, all decisions are synchronous, so the section of the table was obsolete for this exercise. As an example, in the table, the first decision for Queensland Fire and Emergency Services comes from their communication team. They deployed crews to the scene. The boundary objects they used to gain information to make the decision were witnesses, and the decision point was synchronous because, at least in this hypothetical scenario, the crews were successfully contacted, available, and acted on the information to go to the scene as requested. The distribution of the decision point was at an intra-agency level for Queensland Fire and Emergency Services.

This type of analysis would be beneficial to establish teams requiring coordination, linked with the process, tasks, and resources required. Establishing optimal tables would enable comparisons against actual events at incidents. This would prove informative for postincident investigation, as well as for resource allocation and training teams in responder organizations.

Team contextual activity template. The contextual activity template is a representation to show how teams are involved in multiple functions over the totality of the incident. In Table 3, the Contextual Activity Template (adapted from Naikar, Moylan, & Pearce, 2006) for the incident is represented by four situations—evaluation of the incident, rescue and casualty treatment, incident management and investigation, and road clearance and assessment. The different functions required to complete these areas of work at the scene are the initial assessment, emergency response, visual assessment, incident assessment, consultation, road crash rescue, casualty treatment, incident management, road clearance operation, and road assessment. In Table 3, functions are listed in the left column. The occurrence of each function is represented by the box. For example, the initial assessment only occurs during the evaluation phase of the incident; however, incident assessment continues beyond the evaluation phase and across rescue and treatment and incident management and investigation phases.

Ashoori and Burns (2013) extended the Contextual Activity Template to identify team requirements at the scene. In team Contextual Activity Template, team workflow is mapped to represent the distribution of work problems to

TABLE 2: Decision Wheel Table Excerpt for the Incident—Queensland Fire and Emergency Services

	Teams	Task	Boundary Object	Distribution
1	Firecom–crews	Send to scene	Witnesses	Queensland Fire and Emergency Services
2	Firecom–QAScom	Inform/request at scene	Witnesses	Queensland Fire and Emergency Services-Queensland Ambulance Service
3	Crews–Firecom	Visual assessment/ updates	Visual assessment	Queensland Fire and Emergency Services
4	Firecom–Scientific	Send to scene	Visual assessment	Queensland Fire and Emergency Services
5	Firecom–Road crash rescue	Send to scene	Visual assessment	Queensland Fire and Emergency Services
6	Firecom–Environmental Protection Agency	Send to scene	Visual assessment	Queensland Fire and Emergency Services
7	Crews–Queensland Police Service	Advise re HAZMAT	Visual assessment	Queensland Fire and Emergency Services-Queensland Police Service
8	Crews–Queensland Ambulance Service	Advise re HAZMAT	Visual assessment	Queensland Fire and Emergency Services-Queensland Ambulance Service
9	Crews–Traffic Response Unit	Advise re HAZMAT	Visual assessment	Queensland Fire and Emergency Services-Traffic Response Unit
10	Scientific–Firecom	Advise re HAZMAT	Results of testing	Queensland Fire and Emergency Services
11	Scientific–Crews	Advise re HAZMAT	Results of testing	Queensland Fire and Emergency Services
12	RCR–Firecom	Inform re rescue	Casualty	Queensland Fire and Emergency Services

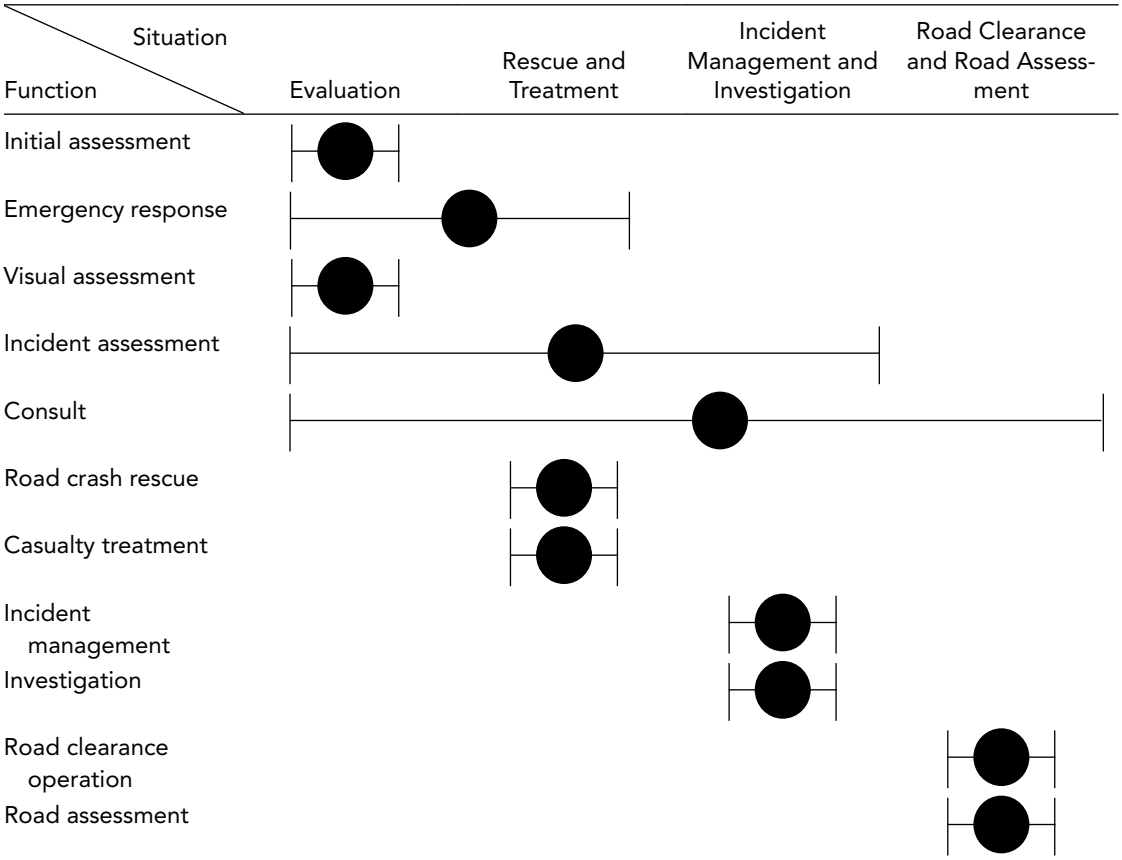
Note. HAZMAT = hazardous material; RCR = road crash rescue.

the different teams. Roles and responsibilities of teams are in the rows and the work problems allocated to teams are illustrated with circles. Communication and coordination associated with reallocation of work problems are shown with arrows. In Figure 7, the work functions of the agency communication teams, general responders, and specialist teams are identified. Due to the complexity of the workflow, the work functions have been coded:

1. IA = Initial assessment
2. ER = Emergency response
3. VA = Visual assessment
4. In E = Incident evaluation
5. C = Consulting
6. ER2 = Second emergency response (representing the sending of specialist teams)
7. CT = Casualty treatment
8. RCR = Road crash rescue
9. IM = Incident management
10. Inv = Investigation
11. RCl = Road clearance
12. RA = Road assessment

As shown by the arrow directions, the horizontal axis in Figure 7 does not specifically represent time.

TABLE 3: Contextual Activity Template of the Incident



Source. Adapted from Naikar, Moylan, and Pearce (2006).

Team Control Task Analysis Findings

In this study, team Control Task Analysis identified areas that shared intra- and interagency elements across the work domain.

The Decision Wheels and accompanying Decision Wheel table clearly outlined the complexity of the traffic incident scene. The hypothetical incident could not be accurately mapped by the wheels, but the exercise identified an optimal scenario between agencies. The use of this tool to map actual incidents, especially when issues occur, is evident.

The basic Contextual Activity Template of the incident identified "consulting" as the only activity required to be conducted across all stages of the incident. The importance of consulting within and between agencies at the scene supports the findings of the current research. In previous research conducted by the authors (Cattermole, Horberry, & Hassall, 2016; Cattermole et al., 2015), a major focus of participants when asked about inefficiencies and safety concerns was the lack of shared understanding and coordination between agencies.

The modified Contextual Activity Template enabled work functions to be grouped according to activities. In accordance with the Work Domain Analysis, this analysis highlighted that all teams have an intra- and interagency requirement for effective incident evaluation (initial response, visual assessment, accident assessment). It is likely that of all functions at the incident, incident evaluation requires the highest level of communication, and shared and complimentary policies and practices. In the study, participants discussed that an interoperability channel is currently being trialed by the Queensland Police Service, Queensland Fire and Emergency

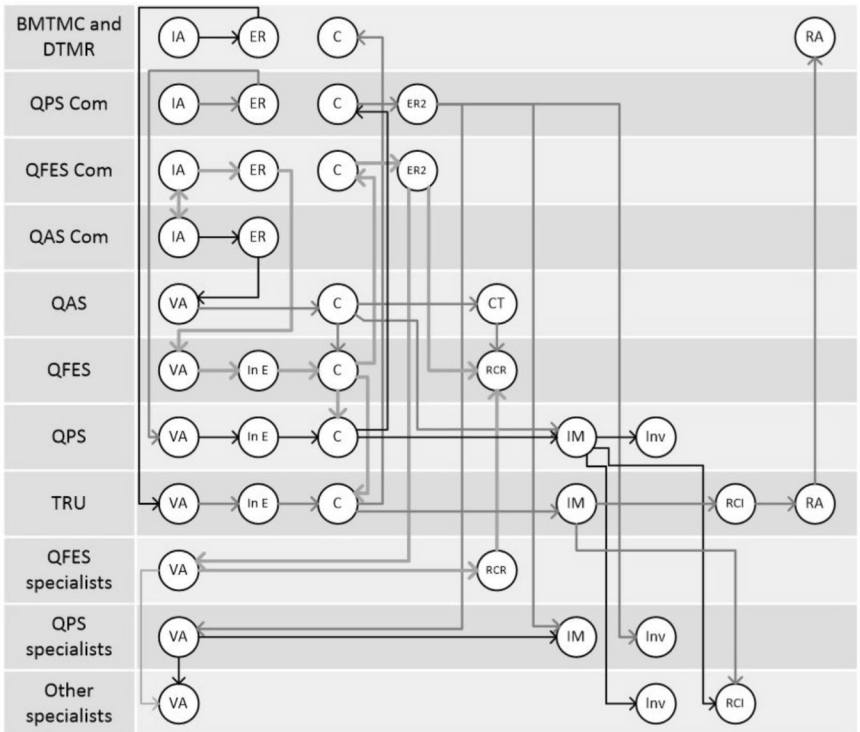


Figure 7. Modified contextual activity template representing the distribution of work functions at the incident.

Note. BMTMC = Brisbane Metropolitan Traffic Management Centre; DTMR = Department of Transport and Main Roads; IA = initial assessment; ER = emergency response; C = consulting; QPS = Queensland Police Service; ER2 = second emergency response; QFES = Queensland Fire and Emergency Services; QAS = Queensland Ambulance Service; CT = casualty treatment; In E = incident evaluation; RCR = road crash rescue; IM = incident management; Inv = investigation; TRU = Traffic Response Unit; VA = visual assessment; RCl = road clearance; RA = road assessment.

Services, and the Queensland Ambulance Service, which may be a reflection of agency understanding of the importance of interagency communication for successful interoperability. However, currently, the channel will only be used between the Queensland Police Service, Queensland Fire and Emergency Services, and Queensland Ambulance Service. This analysis suggests that consideration should perhaps be given to extending the functions so that communication is tailored to the requirements of specific scenes. For example, a communication channel can accept other groups such as the BMTMC, Traffic Response Unit, and specialist groups if the incident requires them to be working in the environment. Also, incident channels should be available so that everyone working at a particular incident can be in contact.

Overall, the Work Domain Analysis and Control Task Analysis sections of the analysis identified shared aspects of the incident and highlighted aspects of the scene that required coordination as well as identifying scene tasks that were individual functions of agencies. This information is useful in determining focus of effort for interagency coordination.

Team Function: Visual Assessment

Factors / Situation	Team structure	Resource access	Expertise level	Task priority	Procedures	Location	Duration	Systems used
HAZMAT	QFES	Full access to QFES only	Expert	Urgent	In case of HAZMAT QFES take incident command and use guidelines to establish the level of emergency. For example, in the current incident it is likely that they would request a 500m evacuation zone, including all QPS and TRU officers.	Incident scene	Quick response is required and QFES remain in command until HAZMAT cleared.	Communication is through Firecomm and procedure is through directives and guidelines.
NORMAL	QFES QPS QAS TRU	Full access	Range of novices to experts	High	Each agency sends teams to complete a visual assessment of the incident according to agency priorities and processes. Teams send information back to their agency.	Incident scene	Priority codes for attending incidents differ according to agency.	Agencies use their own communication systems, procedures and protocols.

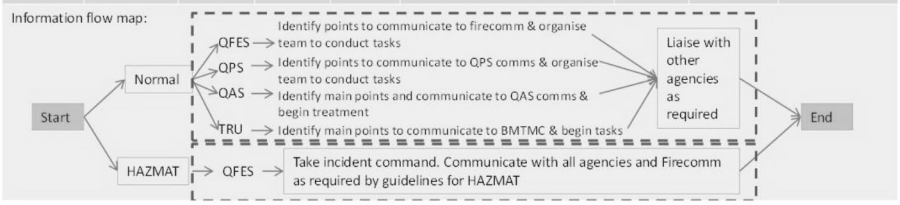


Figure 8. Operational strategies table for visual assessment at traffic incidents.
Note. HAZMAT = hazardous material; QFES = Queensland Fire and Emergency Services; QPS = Queensland Police Service; TRU = Traffic Response Unit; QAS = Queensland Ambulance Service; BMTMC = Brisbane Metropolitan Traffic Management Centre.

Team Strategies Analysis

In team Strategies Analysis, how tasks are executed at incidents is investigated across four categories. The first is operational—strategies that explain how to carry out control tasks. The second category is coordination—strategies for analyzing coordination structures and the process of coordinating structures. The third is team development—strategies that use Tuckman's team development model to understand how behaviors change during the team life cycle (Tuckman & Jensen, 1977). The final category is structural—the strategies that build on work domain constraints (Ashoori et al., 2014). For this exercise, a discussion between responders about the different strategies for incidents involving HAZMAT versus "normal" (non-HAZMAT) incidents enabled an analysis of likely operational strategies and team coordination strategies. Team development strategies analysis and structural strategies require an analysis of actual teams working together, and as this analysis was conducted on a hypothetical incident, it was impossible for these analyses to be conducted.

Operational strategies. Using a modification of Information Flow Maps and the Contextual Activity Template, operational strategies at incident scenes can be identified. In Queensland, regulations dictate that the Queensland Police Service manages the scene of standard traffic incidents. However, in the case of incidents involving HAZMAT, Queensland Fire and Emergency Services become the controlling agency at the scene. In Figure 8, the operational strategy for standard visual assessments by responder agencies is compared with the operational strategy for incidents where there is HAZMAT. This information is useful for traffic incident management process development and also training as it identifies team structures and interaction patterns under different circumstances.

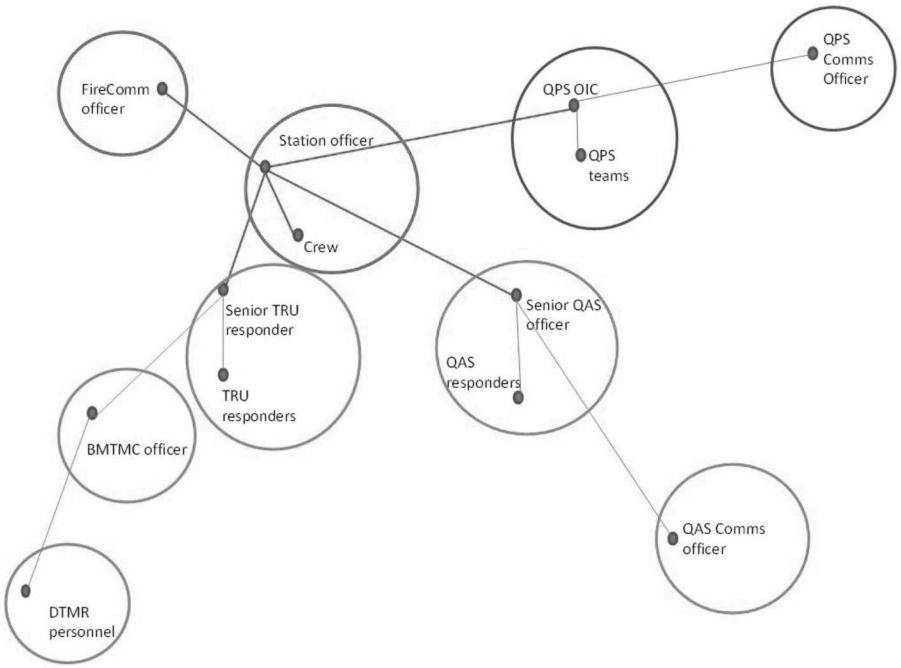


Figure 9. Team coordination strategy for communication of HAZMAT requirements at the incident.
Note. HAZMAT = hazardous material; QPS OIC = Queensland Police Service Office of the Information Commissioner, TRU = Traffic Response Unit; QAS = Queensland Ambulance Service; BMTMC = Brisbane Metropolitan Traffic Management Centre; DTMR = Department of Transport and Main Roads.

Coordination strategies. Using the HAZMAT example from operational strategies, an example coordination strategy can be investigated (Figure 9). Although a HAZMAT incident has a more streamlined operational strategy, communication between agencies and their communication teams remains separate. According to the coordination strategy for communication of HAZMAT requirements following the Queensland Fire and Emergency Services visual assessment, a diplomatic coordination structure is evident. Rasmussen, Pejtersen, & Schmidt (1990) identified that in diplomatic coordination structures, the individual decision makers can only coordinate with their neighbor decision makers and the information traffic is locally planned. In Figure 9, when the station officer of the crew sent to visually assess the incident saw the HAZMAT situation, he or she consults with guidelines and then communicate them to Firecom, the Queensland Police Service officer in charge, the senior Traffic Response Unit responder, and the senior Queensland Ambulance Service officer at the scene. Those officers then contact their own communication teams as well as the lower level officers at the scene. As this is a hypothetical incident, it is impossible to investigate actual coordination processes, but team coordination strategy analysis for traffic incident management could be used to understand the processes underlying effective coordination and to identify poor coordination structures. It is likely, given the identified nonoptimal coordination strategy for this desktop exercise, that the analysis would yield useful results.

Team Strategies Analysis findings

Strategies Analysis has been successfully used to map the complexity of the road system previously. Cornelissen, Salmon, McClure, and Stanton (2012) used a Strategies Analysis diagram to

map different road user types at complex intersections. Ashoori et al. (2014) used team Strategies Analysis to identify different options for actors in medical teams in different workplace situations. For this hypothetical incident, an example was used to test the applicability of the analysis to traffic incident management. Operational strategies for a normal incident versus a HAZMAT incident were quite different. The HAZMAT incident was far more streamlined. This type of analysis would be useful in the traffic incident management environment to map real versus ideal operational strategies of incidents. For example, the participants in this exercise identified that the rules for HAZMAT coordination are strictly followed in the urban environment but less so in the regional/rural environment, largely due to the reduced resources and officer experience in regional/rural areas.

Using the same example, an analysis of coordination strategies for a HAZMAT incident identified that a diplomatic coordination strategy is used in the traffic incident management environment. This strategy is not optimal for HAZMAT incidents due to the urgency of the information flow requirements. Consideration may need to be given to developing alternative coordination strategies for emergency scenarios.

DISCUSSION

This study investigated intra- and interagency team requirements at traffic incidents using a modification of team Cognitive Work Analysis. The aim was to establish collaboration points and to develop recommendations to improve intra- and interagency coordination, collaboration, and interoperability at traffic incident scenes. In line with this aim, the analysis uncovered themes for effective traffic incident management team coordination and generated new knowledge, especially regarding agency interactions, by analyzing traffic incident management holistically.

Improving Traffic Incident Management

Overall, a number of key themes were identified in the study that could be used to focus efforts for emergency responder agencies looking to improve traffic incident management.

The first theme relates to priorities and shared awareness. The lack of shared awareness at the incident scene creates the risk of interoperability issues. As an example, when a fatality occurs at a HAZMAT incident, the police require the scene to remain untouched until an investigation can be carried out. However, the road authority requires the HAZMAT material to be cleared from the road within a time period to prevent damage to the road. The responders from each agency do not have visibility of the competing priorities while at the scene. Therefore, it is likely that one agency will hamper the requirements of the other. If the road authority clears the HAZMAT material, the investigation scene will be corrupted. If the HAZMAT material remains on the road, the road will need to remain closed for days to repair, causing wider community issues. These types of issues indicate the requirement for an overarching governance arrangement between agencies, as well as risk matrices/decision-making processes for complex incidents. In a future study, it would be useful to conduct workshops using human factors and human-centered design techniques to co-design traffic incident management training and processes to better support shared awareness at incidents.

Communication and consultation was another emerging theme that requires consideration. Agency communication teams are not aligned, despite the fact that all agencies share the same communication functions and priorities. Consultation is also a shared requirement by agencies throughout the entirety of an incident. It is likely that aligning communication teams and creating a high-level interagency governance structure for traffic incident management would accommodate this shared requirement.

The desktop exercise was of a complex incident. A requirement to review the current coordination strategy for HAZMAT incidents was evident in the findings of the analysis, but it is likely that all complex incident types need review. Policies and processes written by agencies in isolation despite the fact that the functions at the scene require coordination and collaboration are more likely to become a safety issue in complex environments.

Limitations and Further Research

A limitation of this study is that the scenario was a desktop exercise rather than a real incident. It is possible that processes identified in an office would not occur in real incident situations. A further limitation was that only one incident was analyzed so the team coordination and collaboration requirements identified for the incident might be specific to the incident rather than generalized requirements. A third limitation was that the group were all officers from an urban center. Differences between urban and regional/rural responses were discussed in the exercise and it is likely the analysis could not be generalized to regional/rural incidents. The study could also have benefited from representation from other agencies at the scene—the Queensland Ambulance Service, the Environmental Protection Agency, the Coroner and tow truck operators. Due to the level of experience of the participants, however, they were able to report on likely perspectives and actions of the agencies that were not represented.

Despite the limitations of the study, the potential of team Cognitive Work Analysis as a tool to analyze the traffic incident management environment is evident. Future analyses could conduct studies interviewing decision makers from attending agencies at real incidents, in urban, regional, and rural environments. A study conducted at real incidents would also include asynchronous decisions, which were a noted limitation in team Control Task Analysis. An interesting future study could also include a wider array of officer types from the participating agencies. For example, communication teams, training staff, policy makers, and specialist teams aside from the Queensland Police's Forensic Crash Unit (who were represented) may add extra perspective about the work domain and work flow at traffic incidents. Information from this type of analysis would inform optimal processes and practices in the traffic incident management environment. This paper contributes to the Cognitive Work Analysis (CWA) field by providing an example of applying team CWA to a new domain. The desktop exercise approach used, and the obtained findings about improving interagency coordination and communication, may be relevant to other emergency management fields. Any parallels with other complex environments, for example medical, military, and mining environments, might offer valuable insights to improve traffic incident management. Finally, given the effectiveness of applying team Cognitive Work Analysis to the traffic incident management domain, it would be beneficial for future studies to attempt a similar application across other domains, perhaps requiring some minor adjustments to the methodology as was the case for this study.

CONCLUSION

The traffic incident management system requires intra- and interagency team coordination and collaboration. The current study investigated traffic incident management using an adaptation of team Cognitive Work Analysis. Previously, analyses using team Cognitive Work Analysis were mainly restricted to medical teams. The successful application of team Cognitive Work Analysis into a new domain indicates the framework may be useful more broadly. This analysis was also helpful in raising new issues that may lead to better interagency coordination and communication. Where the individual analyses of previous studies by the authors (Cattermole, Horberry, Burgess-Limerick, Wallis, & Cloete, 2015; Cattermole et al., 2016; Cattermole et al., 2014) identified system issues and solutions related to individuals—for example, the need to create stronger policies to better support novice responders at their first fatality crashes—the group process identified system issues relating to gaps in team coordination, communication and structures as well as the policies and processes related to intra- and interagency teams. The framework enabled an examination of the traffic incident management system as a whole—establishing the purpose, priorities, tasks, and resource requirements at incidents over time. Key weaknesses in the traffic incident management system identified by the analysis were

- disparate priorities at the scene creating the risk of interoperability issues,
- nonalignment of agency communication teams despite the fact that all agencies share the same functions and priorities regarding the tasks related to the communication teams,

- policies written in isolation by separate agencies despite the fact that the functions at the scene require coordination and collaboration between agencies,
- a requirement to improve consultation across agencies and at the incident scene, and
- a requirement to review the current coordination strategy for HAZMAT traffic incidents.

The application of team Cognitive Work Analysis for traffic incident management has produced *novel* insights that could potentially improve the safety of responders working at traffic incidents, and also the effectiveness of the traffic incident management system, leading to improvements to casualty outcomes and reduced risks for motorists and responders in and around traffic incidents.

ACKNOWLEDGMENTS

The authors thank the participants who took place in this research. In addition, they thank Mr. Brendan Lawrence for his research assistance in creating and revising the figures in this paper.

ORCID iD

Tim Horberry https://orcid.org/0000-0002-3453-0216

REFERENCES

Ashoori, M., & Burns, C. (2013). Team cognitive work analysis structure and control tasks. *Journal of Cognitive Engineering and Decision Making*, *7*, 123–140.

Ashoori, M., Burns, C. M., d'Entremont, B., & Momtahan, K. (2014). Using team cognitive work analysis to reveal healthcare team interactions in a birthing unit. *Ergonomics*, *57*, 973–986.

Carson, J.L. (2010). *Best Practices in traffic incident management*. Federal Highway Administration Report # FHWA-HOP-10-050.

Cattermole, V. T., Horberry, T., Burgess-Limerick, R., Wallis, G., & Cloete, S. (2015, October). *Using the critical decision method and decision ladders to analyse traffic incident management system issues*. Paper presented at the 1st Australasian Road Safety Conference, Gold Coast, Australia.

Cattermole, V. T., Horberry, T., & Hassall, M. (2016). Using naturalistic decision making to identify support requirements in the traffic incident management work environment. *Journal of Cognitive Engineering and Decision Making*, *10*, 309–324.

Cattermole, V. T., Horberry, T., Wallis, G., & Cloete, S. (2014, November 12–14). *An operator-centred investigation of safety issues for emergency responders at traffic incidents*. Paper presented at the Australasian Road Safety Research, Policing and Education Conference, Melbourne, Australia.

Cattermole-Terzic, V. (2017). *A human factors investigation into the effectiveness of traffic incident management systems* (Unpublished PhD thesis). The University of Queensland, Brisbane, Australia.

Charles, P. (2007a). *Review of current traffic incident management practices* (Research Report No. R297/07). Sydney, Australia: Austroads.

Charles, P. (2007b). *Traffic incident management guide to best practices* (Research Report No. AP-R298/07). Sydney, Australia: Austroads.

Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J. L. (2013). Interactive team cognition. *Cognitive Science*, *37*, 255–285.

Cornelissen, M., Salmon, P. M., McClure, R., & Stanton, N. A. (2013). Using cognitive work analysis and the strategies analysis diagram to understand variability in road user behaviour at intersections. *Ergonomics*, *56*(5), 764–780.

Farradyne, P. B. (2000). *Traffic incident management handbook*. Prepared for Federal Highway Administration, Office of Travel Management. Retrieved from https://hcm2010.org/system/datas/65/original/Traffic%20Incident%20Management%20Handbook.pdf

Fiore, S. M., & Salas, E. (2004). Why we need team cognition. In E. Salas & S. M. Fiore (Eds.), *Team cognition: Understanding the factors that drive process and performance* (pp. 235–248). Washington, DC: American Psychological Association.

Fischer, J., Krzmarzick, A., Menon, A., & Shankwitz, C. (2012). *Performance analysis of squad car lighting, retro-reflective markings, and paint treatments to improve safety at roadside traffic stops* (Report CTS No. 12–13). Minneapolis: University of Minnesota.

Flavell, J. H., & Wellman, H. M. (1975, August). *Metamemory*. Paper presented at the 83rd Annual Meeting of the American Psychological Association, Chicago, IL.

Gutwin, C., & Greenberg, S. (2002). A descriptive framework of workspace awareness for real-time groupware. *Computer Supported Cooperative Work*, *11*, 411–446.

Kenny, D. A., & La Voie, L. (1984). The social relations model. *Advances in Experimental Social Psychology*, *18*, 141–182.

Klein, G. (1998). *Sources of power: How people make decisions*. Cambridge, UK: The MIT Press.

Klein, G., Calderwood, R., & Macgregor, D. (1989). Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics*, *19*, 462–472.

Klein, G., & Wright, C. (2016). Macrocognition: From theory to toolbox. *Frontiers in Psychology*, *7*, Article 54.

Langan-Fox, J., Code, S., & Langfield-Smith, K. (2000). Team mental models: Techniques, methods, and analytic approaches. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *42*, 242–271.

McNeese, M. D., Rentsch, J. R., & Perusich, K. (2000). Modeling, measuring, and mediating teamwork: The use of fuzzy cognitive maps and team member schema similarity to enhance BMC/sup 3/I decision making. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, *2*, 1081–1086.

Naikar, N., & Elix, B. (2016). Reflections on cognitive work analysis and its capacity to support designing for adaptation. *Journal of Cognitive Engineering and Decision Making*, *10*, 123–125.

Naikar, N., Moylan, A., & Pearce, B. (2006). Analysing activity in complex systems with cognitive work analysis: Concepts, guidelines and case study for control task analysis. *Theoretical Issues in Ergonomics Science*, *7*, 371–394.

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance

models. *IEEE Transactions on Systems, Man, and Cybernetics*, *13*, 257–266.

Rasmussen, J., Pejtersen, A. M., & Schmidt, K. (1990). *Taxonomy for cognitive work analysis*. Roskilde, Denmark: Risø National Laboratory.

Rentsch, J. R., Mello, A. L., & Delise, L. A. (2010). Collaboration and meaning analysis process in intense problem solving teams. *Theoretical Issues in Ergonomics Science*, *11*, 287–303.

Salas, E. E., & Fiore, S. M. (2004). *Team cognition: Understanding the factors that drive process and performance*. Washington, DC: American Psychological Association.

Scherer, R. F., & Petrick, J. A. (2001). The effects of gender role orientation on team schema: A multivariate analysis of indicators in a US federal health care organization. *The Journal of Social Psychology*, *141*, 7–22.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*, 127–190.

Tuckman, B. W., & Jensen, M. A. C. (1977). Stages of small-group development revisited. *Group & Organization Management*, *2*(4), 419–427.

Vicente, K. J. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. Mahwah, NJ. Lawrence Erlbaum.

Vanessa Cattermole-Terzic is the Director, Research and Insights for Queensland's Department of Transport and Main Roads, with 20 years' experience in government roles across transport, police and education agencies.

Tim Horberry works at both Monash University Accident Research Centre in Australia and Coventry University in the UK. He has conducted applied human factors research in the road transport, mining, rail and medical domains for over 25 years.

Transparency in Autonomous Teammates: Intention to Support as Teaming Information

April Rose Panganiban, Wright-Patterson Air Force Base, USA, **Gerald Matthews**, University of Central Florida, USA, and **Michael D. Long**, Wright-Patterson Air Force Base, USA

Human–Machine teaming is a very near term standard for many occupational settings and still requires considerations for the design of autonomous teammates (ATs). Transparency of system processes is important for human–machine interaction and reliance but standards for its implementation are still being explored. Embedding social cues is a potential design approach, which may capture the social benefits of a team environment, yet vary with task setting. The current study examined the manipulation of transparency of benevolent intent from an AT within a piloting task requiring suppression of enemy defenses. Specifically, the benevolent AT maintained task communication as in a neutral condition, but included messages of support and awareness of errors. Benevolent communication reduced reported workload and increased reported team collaboration, indicating that this team intent was beneficial. In addition, trust and acceptance of the AT were rated higher by individuals tasked with depending on the system to protect them from missile threats. The need for information from ATs is beneficial, however may vary depending on team type.

Keywords: human–machine teaming, human–robot teaming, laboratory study, military

Address correspondence to April Rose Panganiban, Air Force Research Laboratory, Wright-Patterson Air Force Base, 711th Human Performance Wing, Human Trust and Interaction Branch, Wright-Patterson Air Force Base, OH 45433, april_rose.panganiban@us.af.mil.

Journal of Cognitive Engineering and Decision Making
2020, Volume 14, Number 2, June 2020, pp. 174–190
DOI: 10.1177/1555343419881563

The advancement of technology using advanced software algorithms, machine learning, and coordination of sensors has evolved automated system abilities toward autonomous functioning. Autonomy differs from conventional automated systems in the ability of the system to monitor its state and make decisions on its own (Barnes, Chen, & Hill, 2017). The presence of an autonomous teammate (AT) has many benefits to a human operator in that machine capabilities of the AT such as processing speed and fewer physical limitations can enhance joint decision-making and action. Allocation of tasks to the AT may increase the human's functional bandwidth. This transition in technical capability brings forth many research questions in the area of Human–Machine Teams (HMTs).

Activities that occur in a team as they work toward a common goal involve several, ongoing, stages of planning and monitoring as well as interpersonal processes (LePine, Piccolo, Jackson, Mathieu, & Saul, 2008; Marks, Mathieu, & Zaccaro, 2001). Prior to HMT, humans did not interact with machines in such continuous cooperation. The human conducted tasks in the world with the automated system serving as another source of data or as another output channel (see Figure 1). Issues in human performance were limited to errors in interpretation of the data, maintaining situation awareness (SA) or permitting flawed outputs. As machines have evolved in their capability to receive information (monitor), process and plan, and interact with the human, the system has become more like a teammate than a tool. In this relationship, the human has less direct insight into the actions and processes of the machine and their immediate impact on the shared environment (see Figure 2). Thus, contemporary HMT requires the examination of how interacting with an AT may influence trust, stress, and workload. In human–human teams, the relationships

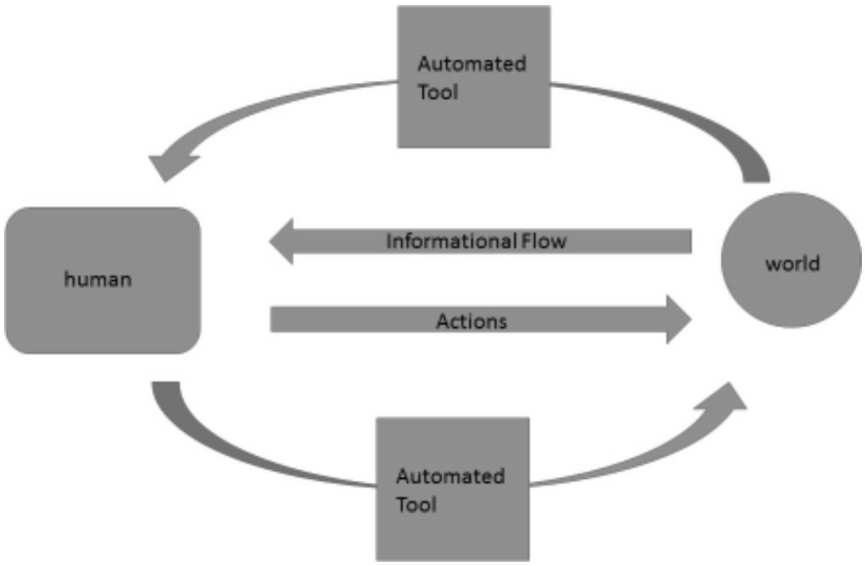


Figure 1. Human decision-making with low-level automation.
Note. Information is transferred unaltered to or from the human.

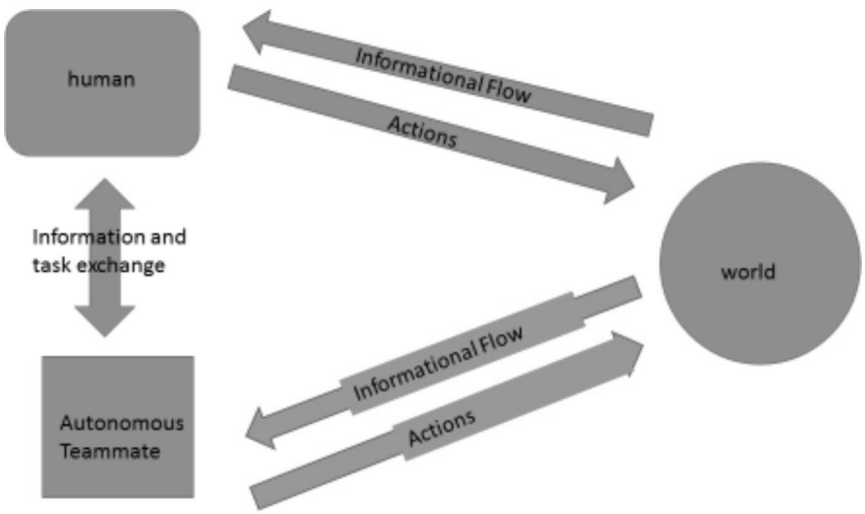


Figure 2. Human–robot teaming.
Note. Note that information fed to the system is filtered rapidly and outputted either directly to the world or in altered form to the human.

between trust and communication on performance (De Jong, Dirks, & Gillespie, 2012; Marlow, Lacerenza, Paoletti, Burke, & Salas, 2018) have been given appropriate attention. Much is known about how teams interact and the various processes they carry out to achieve their goals (LePine et al., 2008; Marks et al., 2001). However, teamwork processes in HMTs are unfamiliar territory, raising the important question of "How should HMTs interact to maintain trust, ensure reliance, and reduce stress?"

Design suggestions speak to social adaptations for a machine or system partner. For example, reducing interruptions can be seen as teammate

"etiquette" between adaptive automation and the human via reduced messaging (Dorneich, Ververs, Mathan, Whitlow, & Hayes, 2012). However, the suggestion fails to capture the dynamics of the teamwork process. Other approaches toward design of HMT point toward other social constructs like "empathy" or informational features such as "transparency" (Leite et al., 2013; Lyons, 2013). These features are naturally revealed in the interaction between human–human teams through communication between team members and may best be implemented together in the design of a machine partner (Klein, Woods, Bradshaw, Hoffman, & Feltovich, 2004).

The purpose of this study was to investigate how ATs might be designed to support HMT in military aviation, focusing on the social interaction between the human and the AT, specifically machine-to-human communication. We investigated multiple outcomes that can be used to evaluate the effectiveness of HMT, including team processes, stress and workload, trust, and confidence. We treated team perceptions and trust as multifaceted constructs. Critical aspects of team functioning identified in previous work include coordination and cohesion (Marks et al., 2001). Trust can be evaluated through scales derived from both human- and machine-oriented studies (e.g., Madsen & Gregor, 2000), as well as technology acceptance (Ghazizadeh, Peng, Lee, & Boyle, 2012).

In the remainder of this introduction, we first discuss evidence from human–human team research that affect management is important for team cohesion and for optimizing trust and team performance. We then discuss evidence from social robotics that affect regulation is similarly important for HMT, especially when the machine can function autonomously. Next, we introduce transparency as a design feature that can support affect regulation by signaling machine benevolence. The present study aimed to test whether transparency has the benefits suggested by the research reviewed, in the context of a military air mission.

HUMAN–HUMAN TEAMING

Human teams are groups of at least two individuals, working toward a common goal. In occupational settings, teaming effectiveness arises from the successful management of both explicit task work and teamwork that supports attainment of task goals (Marks et al., 2001). Teamwork has many subprocesses, which involve coordinating, monitoring, planning, and affect management. Marks et al. (2001) noted that coordination can occur in active and inactive phases. Inactive or transition phases involve mission analysis, planning and strategizing, and goal specification where feedback can occur between members. During action or active phases, teamwork is characterized by various coordinated monitoring and backup behaviors (Marks et al., 2001). These activities organize the procedures required for team actions, typically involving feedback or provision of assistance to other team members. Finally, several social processes are necessary for teamwork and task work in the form of conflict management, motivation, and affect management.

Interpersonal processes that serve to regulate affect can contribute positively to both teamwork and task work (Marks et al., 2001). Actions supporting task work may be missed, delayed, or carried out incorrectly if a team member's affective state is suboptimal. Stress can result both from the cognitive demands of task work (Matthews, Wohleber, & Lin, 2019) and from maladaptive interpersonal processes such as conflict (Ilies, Johnson, Judge, & Keeney, 2011). Stress, if not mitigated by adaptive coping, reduces teammate support through attentional narrowing and inattention to others or simply increased insensitivity to others (Driskell, Salas, & Driskell, 2018). Regulation of one another's state is another important teaming process, which can prevent the stress-related cognitive burden on the individual partner and also reduce negative mood. Negative mood can reduce team awareness through the stressed partner becoming disengaged from the team or shifting attention from the larger team goals to immediate, local stimuli (Pfaff, 2012).

A tactic for adaptively regulating a teammate's mood is to communicate a willingness to be supportive, which signifies benevolence toward the teammate. Benevolence is a trustworthiness characteristic, which increases trust, a willingness to be vulnerable to the unmonitored actions of the object of trust, from the trustor (Mayer,

Davis, & Schoorman, 1995). Trust in a teammate may increase reliance on a teammate; therefore, trust can increase the frequency and acceptance of team backup behaviors (Smith-Jentsch, Kraiger, Cannon-Bowers, & Salas, 2009).

It is clear that task coordination and affect management interact to affect trust, stress, and cognitive resources in challenging human–human team settings (Driskell et al., 2018). As machines are built to work with humans as equal teammates, factors that influence affect and support coordination will be important in operator state and performance. Therefore, consideration of the impact of social factors between human teammates derived from team coordination may be just as important in human–machine teaming as they are in human–human teaming. Affective management and support can be seen, as transparency of team intent which is cited as an important factor for trust and in the design of human–machine teaming.

SOCIAL ROBOTICS AND AFFECT MANAGEMENT FOR AUTONOMOUS TEAMING

Research from the world of social robotics can inform the design of emotionally supportive ATs, including other machines and software agents as well as robots. Humans readily socialize with computers and robots, often drawn to their design and interactive styles (see Wynne & Lyons, 2018). Nass, Fogg, and Moon (1996) found that instructing individuals that their outcomes are tied to performance by a computer "partner" influences the human to perceive better information quality, and to feel more cooperative and open to influence by the computer. When individuals are placed in a context that requires teaming, they report stronger feelings of teaming with the computer than individuals whose performance evaluation is based on their work alone. Teaming tendency is so influential that people will even adjust their answers to conform with a computer partner compared with independent performance (Nass et al., 1996).

Studies of human–robot interaction show how proper implementation of supportive teammate qualities may facilitate both teamwork and task work, and consequently team performance. Displayed empathy from a robotic partner can increase prosocial behavior and motivation (Leite et al., 2013), which may increase trust and important teamwork processes. Empathy requires knowledge of the situation and how it may affect another individual. In a study of supportive robotic interaction, empathy increased feelings of companionship, reliable alliance, and self-validation (Leite et al., 2013). In a teaming context, these feelings may be expressed as emergent states of cohesion, trust, and motivation, respectively, supporting improved team performance. Design of a personable AT may also have benefits for team cohesion. Anthropomorphic features of agents contribute to trust resilience (de Visser et al., 2016). In de Visser et al.'s (2016) study, anthropomorphism elevated self-reported trust after an error in a decision-making task, especially under conditions of uncertainty. Human-like behavior purposed to repair trust also increased trust in the automation.

Design of ATs to indicate affective qualities like empathy may not only promote acceptance of robotic partners as teammates, but also serves to provide information to bolster taskwork and situational awareness under complex and time-sensitive circumstances. Having awareness of a robotic partner's intention and state is a needed step toward optimal human–machine teaming because it reduces workload of the human partner in needing to attend to and supervise a robotic partner (Groom & Nass, 2007).

SUPPORTING AFFECT MANAGEMENT WITH TRANSPARENCY

Transparency is the intentional design of a system to communicate its capabilities and current state (Barnes et al., 2017; Lyons, 2013; Selkowitz, Lakhmani, Chen, & Boyce, 2015). It can be used to finely tune a human operator's perception of the AT's ability, intent, and situational constraints (Lyons, 2013). Lyons (2013) classifies transparency into several models of schemas: intentional, task, analytical, environmental, team, and human state. Transparency in the AT context parallels human–human team information in that the intentional schema in AT explains the purpose and design of the AT while it might be represented in human partners

as the role and training of that team member. Transparency of task model information, such as the AT's understanding of the task and how to execute it, supports a team's task knowledge. Shared awareness of the system should result in both good team performance and increased trust. For example, explaining the logic of an emergency landing aid resulted in increased reported trust in commercial airline pilots when compared with sharing confidence of choices or no transparency at all (Lyons et al., 2017).

Transparency of state-based information creates a foundation for true HMT. For example, robotic partners may become aware of their performance limitations given changes in the environment and should communicate this understanding to the human partner to ensure appropriate reliance (Lyons, 2013). In addition, indicated awareness from a robotic partner that the human partner is under distress may have several positive benefits for the team. Such communication provides a cue that reprioritizing of team member tasks should occur. Thus, transparent design of AT interaction may promote perceptions of benevolence from a robotic partner and consequently mitigate stress and negative emotion. Knowing that a robotic system seeks to support the well-being and performance of a human operator may facilitate trust of that system during situations characterized by high uncertainty and stress. Indeed, trust may be linked to adaptive processes associated with positive affect, including coping (Schaefer & Scribner, 2015) and confidence (Hoff & Bashir, 2015; Lyons et al., 2016).

THE PRESENT STUDY

The current study investigated how transparency of the AT's social intent (benevolence) affected perceptions of teaming, workload and stress, trust in the AT, and confidence. The AT was an autonomous wingman accompanying the participant's (i.e., pilot's) plane, equipped with surveillance and missile-jamming capabilities to support a military air mission. We manipulated transparency by contrasting benevolent and neutral conditions. In the former case, the AT communicated its intention to support the human and to correct its errors, signaling its awareness of the human partner's expectations. In the neutral condition, the AT provided only information on its immediate task activities, with no additional transparency into its intentions.

The study also manipulated the type of human–machine partnership by varying level of dependence of the human on the AT. In both independent and dependent teaming conditions, the AT broadly supported the mission. However, in the dependent teaming condition, the AT was specifically tasked with countering the threat to the human posed by enemy Surface-to-Air Missiles (SAMs). This contrasts with the independent teaming condition where the AT only performs surveillance and the human performs the mission and counters the SAM threat.

Consistent with the role of multiple cognitive, social, and affective processes in teaming (Groom & Nass, 2007; Marks et al., 2001; Wynne & Lyons, 2018), there are various metrics for the effectiveness of HMT, which may diverge from one another (e.g., Lyons & Guznov, 2018). Thus, we secured multiple measures to assess four constructs that may signal effective HMT: positive teaming perceptions, low stress and workload, trust in the AT, and confidence. Measures were selected to assess key facets of each construct.

Several hypotheses were made regarding the positive effect of the benevolent transparency manipulation on the various outcome measures. An AT that communicates awareness of a discrepancy in its performance from the human partner's expectation provides a form of transparency, which implies deviation from goal but also signals an intention to correct the deviation, thus supporting the pilot. Thus, this process should have multiple benefits for perceptions of teaming. Specifically, signaling benevolence should enhance team coordination and cohesion, supporting perceived team effectiveness and trust at the team level. Perceptions of teaming were expected to be especially sensitive to benevolence in the dependent condition, in which the human is directly vulnerable to a failure by the AT to neutralize SAMs. Benevolence should also reduce stress related to the task and subjective workload, both of which may increase if the participant is unaware of the AT partner's state and intention. Combined with benevolence,

awareness of a correction should counteract negative perceptions of ability, thus increasing trust and acceptance of the technology. The impacts on trust should be especially salient when performance of the pilot is dependent on actions of the AT. Benevolence may also produce general feelings of confidence in the AT regardless of level of dependence overriding feelings linked to specific system functions (Lee & See, 2004). The following hypotheses were tested in this Air Mission task:

- **Hypothesis 1 (H1):** Benevolent transparency will increase teaming perceptions, specifically in the dependent teaming condition, that is, a Team Type × Transparency interaction was anticipated.
- **Hypothesis 2 (H2):** Benevolent transparency will reduce stress and workload overall, and enhance coping. This effect should be stronger in the dependent teaming condition, that is, a Team Type × Transparency interaction.
- **Hypothesis 3 (H3):** Benevolent transparency will increase multiple facets of trust, including technology acceptance and perceived trustworthiness and trust, that is, a Team Type × Transparency interaction.
- **Hypothesis 4 (H4):** Benevolent transparency will increase confidence in the partner irrespective of dependence, that is, a main effect for transparency.

METHOD

Experimental Design

A 2 × 2 mixed-model design was used where Team Type (Independent/Dependent) was a between-subjects variable and Transparency (Neutral/Benevolent) was a within-subjects variable. The task was a manufactured air-to-ground combat mission where the pilot was tasked with flying a predetermined mission with specific waypoints on approach to a targeted enemy ground missile.

Team Type was defined as the way that the human pilot performed their flight mission with the Viper 2 AT. Participants performed independently of Viper 2 (independent teaming condition) or they were dependent on the action of Viper 2 (dependent condition). Transparency was a counterbalanced within-subject manipulation of Viper 2's communication. In the Neutral condition, language was only informative, while in the Benevolent condition, Viper 2 communicated social support and awareness it had made a heading error.

Participants

Participants (n = 40; 31 males and nine females) were recruited from the local community near a Midwestern military base. Participants did not have flying experience outside of flight simulators or video games. All participants were between the ages of 18 and 40 years with normal hearing and normal or corrected vision.

Participants were assigned at random to either the dependent teaming condition or the independent teaming condition. Both conditions consisted of 12 trials presented in a randomized order, varying in approach directions (heading) toward the goal. All participants were given mission cards that detailed the heading, altitude, and speed that must be maintained by both Viper 2 and the pilot to be safe from SAM detection on approach to the ground missile. Participants attended the lab for a single session of approximately 240 min. Subjective surveys were administered before the task began, between each trial, and at the end of the full set of trials. Each trial lasted 3 min with a total of 12 trials.

Task Design

The task was performed in a multiprojector flight simulator (see Figure 3). Communication was available with the AT, named Viper 2, so that "he" received specific phrases and answered. The voice used was decided in pilot testing where individuals favored the masculine, Texan voice. Each trial consisted of a different set of waypoint headings, or directional changes on the way to the target. The AT was always part of this flight plan, but his role varied with team type. Across all conditions, Viper 2 was expected, through training, to follow the preplanned heading, altitude, and speed. In all conditions, Viper 2 communicated his heading, speed, and altitude at each of the three waypoints on the predetermined flight path. This was meant to establish known team goals. However, as part of the study design, Viper 2 consistently made an error in heading on

Figure 3. Flight mission simulator.

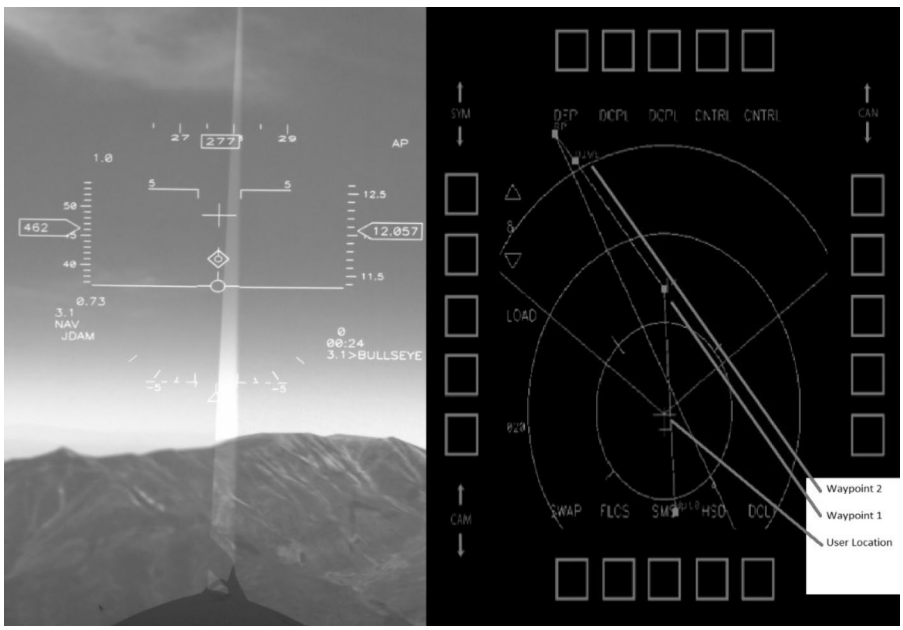


Figure 4. Left: The heads-up display on the screen showing heading, speed, and waypoint distance. Right: The map displayed in the head-down display showing a map with the waypoints and the participant's location.

one of the three waypoints but corrected to fly toward the enemy target. The waypoint in which Viper 2 reported his error was randomized for each trial. Figure 3 shows the experimental setup while Figure 4 shows a zoomed-in view of the main displays for following heading, speed, and waypoint.

Transparency was a counterbalanced within-subject manipulation of Viper 2's communication. In the Neutral condition, language was only informative (i.e., "SAM approaching, radar on," "performing surveillance," "ready for the mission"). When waypoint errors occurred, Viper 2 did not state awareness of them. In the Benevolent condition, Viper 2 communicated social support and awareness he had made a heading error but was correcting (e.g., "I hit that turn a little wrong, correcting," "SAM approaching, radar on; I've got you covered," "Ready to support you"). Social support was meant to

communicate "Benevolent intention," manipulated here as an intent to "keep (the pilot) covered" as well as to inform the operator that he was aware of the flight goals by communicating intent to correct errors.

Team Type was defined as the way that the human pilot performed their flight mission with the Viper 2 AT. In the independent teaming condition, the pilot must attack the missile while cloaking the plane from detection by a defensive SAM near the ground missile. During this task, Viper 2 performed surveillance along the flight path. Thus, human and AT were independent of one another's main task but dependent on each other to accomplish the final mission. In the dependent condition, interdependence was consistent with the pilot carrying out the same mission described. However, Viper 2's purpose was also to use his own onboard device to cloak the pilot from detection by the defensive SAM.

Subjective Measures

Teaming perception. Perceptions of team quality were assessed using a set of items aimed at measuring teamwork-relevant constructs, all using a 7-point-Likert-type scale with scores from each item averaged across relevant items. The constructs consisted of collective efficacy, team collaboration, team trust, and group cohesion. Collective efficacy (based on Riggs & Knight, 1994) was comprised of seven items related to the team's ability to accomplishing objectives (a = .91). Ten items related to team collaboration were developed for this study, specifically related to team communication and how information is exchanged between partners (a = .82). Team trust was based on Naquin and Paulson's (2003) interpersonal trust scale. Items had a focus on both negative trust and positive trust (a = .94). Finally, team group cohesion (based on Rozell & Gundersen, 2003) was assessed with 14 items ascertaining to an individual's enjoyment for and feeling of belonging to their group (a = .94).

Trust. To assess trust in the AT, the Human–Computer Trust (HCT) scale (Madsen & Gregor, 2000) was altered. This scale measures the trust a user has in a computer system using a 5-point Likert-type scale to measure three components of trust: perceived reliability, perceived technical competence, and perceived understandability. The scale consisted of statements such as "Viper 2 performs reliably" or "It is easy to follow what Viper 2 does." Items were averaged within each component maintaining internal consistency for perceived reliability (a = .86), perceived technical competence (a = .85), and perceived understandability (a = .88).

The Trust and Trustworthiness and Propensity to Trust (TTP: Mayer et al., 1995) scale was adapted to assess only the subject's perceptions of ability, benevolence, and integrity of the AT. Items were altered to address Viper 2 versus top management. Ability addressed Viper 2's capability; benevolence was aimed at Viper 2's intention to support; and integrity addressed Viper 2's consistency in acting. All items were assessed on a 5-point Likert-type scale where 1 = *disagree strongly* and 5 = *agree strongly*. Scores for each factor were obtained by averaging across related items.

The Technology Acceptance Measure (TAM) was a mix of items measuring five components: competence, integrity, usefulness, trust, and intention to adopt. These items were adapted from Wang and Benbasat (2005), Komiak and Benbasat (2006), Naquin and Paulson (2003), Li, Hess, and Valacich (2008), and Cheung and Lee (2001) changing phrases to reference a "partner" and the "mission." The present study's version of the survey used a 7-point Likert-type scale with 1 = *strongly disagree* and 7 = *strongly agree*. Current items continued to maintain internal consistency across competence (a = .93), integrity (a = .77), usefulness (a = .90), Trust (a = .83), and intention to adopt (a = .89). Scores for each construct were obtained by averaging related items.

Stress and workload. The Dundee Stress State Questionnaire (DSSQ: Matthews et al., 2002) was administered pre and posttask. Stress state was assessed as variance in task engagement, distress, and worry. The DSSQ factors are calculated as weighted sums of standardized scale scores.

Coping style was measured with the Coping Inventory for Task Situations (CITS: Matthews & Campbell, 1998). It has items relating to task-focused, emotion-focused, and avoidance

coping. Responses were assessed on a 5-point scale where 0 = *not at all* and 4 = *extremely*. Scores on the coping items were summed to find a total score for each of the three scales (no weights).

Individual perceptions of workload were obtained using the NASA Task Load Index (TLX: Hart & Staveland, 1988). The six subscales were answered on a continuous scale from 0 to 100 (0 = *low* and 100 = *high*) for perceived mental demand, temporal demand, effort, frustration, performance, and physical demand. Performance was reverse-scaled such that 0 = *high workload* and 100 = *low workload*.

Decision-making confidence. Individuals were asked about their confidence in their own decision-making and that of their autonomous partners. After each trial, participants reported their confidence on a 100-point, continuous scale (0 = *least confident* and 100 = *most confident*). The two questions asked in the survey were the following: "How confident are you in your decision to continue or abort the last mission?" and "How confident were you in your teammate during the last mission?"

Procedure

All participants, upon arrival, read and signed the informed consent document. They then filled out the pretask DSSQ and some demographic questions.

Participants performed six trials in the Benevolent condition, and six in the Neutral condition, in counterbalanced order. Participants filled out the decision-making confidence scale following every trial. Once participants completed the first block of six trials (either the Benevolent or the Neutral block), they responded to the posttask DSSQ, the team perceptions survey, the NASA TLX, the HCT, the TTP, and TAM. They then completed the second block of six trials and completed the same set of surveys again.

After participants chose the Call Sign for Viper 2 to refer to them by, they were trained first on the controls and the display for the task, and also how to shoot the targets. In the second practice trial, participants were taught how to communicate with Viper 2, and were run through all the possible commands that can be given and messages received by Viper 2. After the third waypoint in each trial, participants are asked by Viper 2 "(Call sign) do you wish to continue with the mission?" Participants could choose to either continue or abort the mission. For the practice trials, participants were told to continue the mission, but were told that for the experimental trials they can abort if they feel the mission will not be successful. On the third practice trial, after saying they would continue the mission, participants experienced being shot down. They were informed after the final practice that for the experimental trials, they would not be shooting the targets, but would only have to say if they wanted to continue or abort, and the trial would end. They were also told that the practice trials were designed so that they could experience events following the decision to continue or not, including the possibility of being shot down, success in Practice Trial 2, and failure in Practice Trial 3. All participants were encouraged to rely on their "gut" feeling about continuing or aborting the mission based on their experience during the practice trial, being aware that they could be shot down. They received no feedback about if they made the correct decision or not. This was done to ensure that participants focused on the interaction with Viper 2, by paying attention to waypoint information as instructed as opposed to creating and additional reason for the outcome. Also by maintaining feelings of risk in the scenario, participants would be more inclined to consider their subjective feelings of trust, teaming, and stress.

RESULTS

Missing data comprised less than 10% of total data and averages were used as replacement for values. The absence of data was due to computer errors in logging or due to experimenter error (e.g., participant closing a survey or experimenter failing to open the survey). However, if participants were missing data from multiple trials (e.g., pretask and posttask DSSQ), they were removed from analyses. The absence of these participant's data are reflected in the specific analyses' reported degrees of freedom. All data were screened for outliers, defined as data having a $z_{\text{score}} > |2.58|$, which

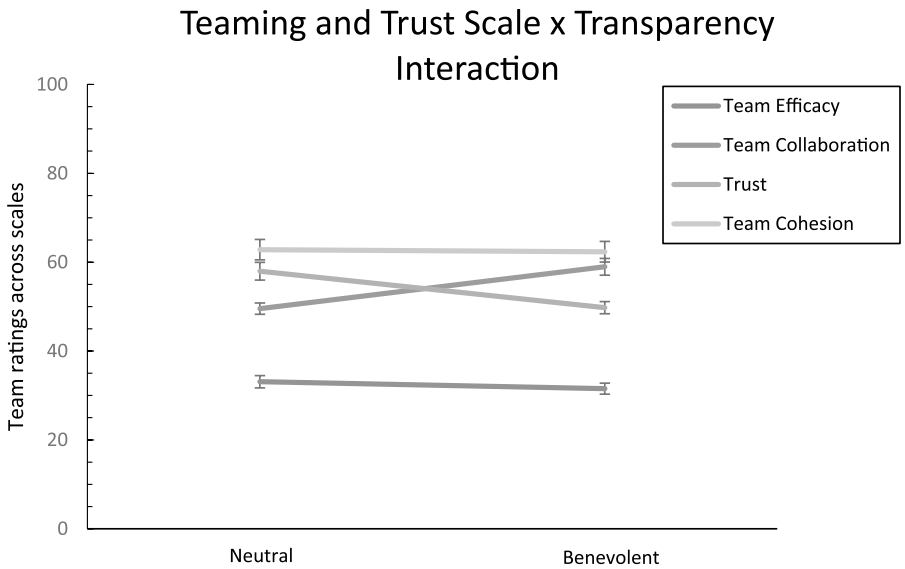


Figure 5. Interaction between transparency and the scales for teaming and trust.
Note. Team collaboration and trust significantly differed with transparency condition ($p < .001$). Standard error bars shown.

were replaced with the next largest value. Data were screened for skewness and kurtosis, which revealed that all data sets were approximately normally distributed (skewness $<|2|$ and kurtosis $<|2|$). The Greenhouse–Geisser correction was applied in all analyses where sphericity was violated. Analyses of variance (ANOVAs) were supplemented with post hoc *t* tests to compare means. These tests were Bonferroni-corrected; the number of comparisons and corrected *p* levels for significance are provided below where appropriate.

Teaming and Trust Questionnaire

To test H1, a 2 (Team Type: Independent/Dependent) × 2 (Transparency: Neutral/Benevolent) × 4 (Scale: Team Efficacy/Team Collaboration/Trust/Team Cohesion) mixed ANOVA was performed. There was a significant interaction between Transparency and Scale, $F(1.89, 53.76) = 30.91$, $p < .001$, $\eta_p^2 = .47$ (see Figure 5). Paired *t* tests (four comparisons; critical $p = .013$) revealed that Transparency had differential effects on measures of team collaboration ($p < .001$). Team collaboration was rated as higher in the Benevolent condition ($M = 58.96$) than the Neutral condition ($M = 49.55$).

Stress and Workload

To test H2, a 2 × 2 mixed-model analysis of covariance (ANCOVA) was performed on each of the DSSQ scales of worry, task engagement, and distress with Team Type (Independent/Dependent) and Transparency (Neutral/Benevolent) as independent variables and pretask scores served as a covariate. The analyses revealed no significant main effects or interactions for the DSSQ scales.

An additional 2 (Team Type: Independent/Dependent) × 2 (Transparency: Neutral/Benevolent) mixed-model ANOVA was run for each of the three coping scales. No significant main effects or interactions were found; however, task-focused coping showed an interaction trend, $F(1, 38) = 3.89$, $p = .056$, $\eta_p^2 = .093$, between Team Type and Transparency shown in Figure 6.

A 2 (Team Type: Independent/Dependent) × 2 (Transparency: Neutral/Benevolent) × 6 (Scale: Mental Demand/Effort/Frustration/Physical Demand/Temporal Demand/Performance) mixed-model ANOVA was run for the NASA TLX. There was a significant main effect for transparency, $F(1, 39) = 5.245$, $p < .05$, $\eta_p^2 = .12$ (see Figure 7). In the benevolent condition, participants reported lower perceived workload ($M = 34.60$) than in the

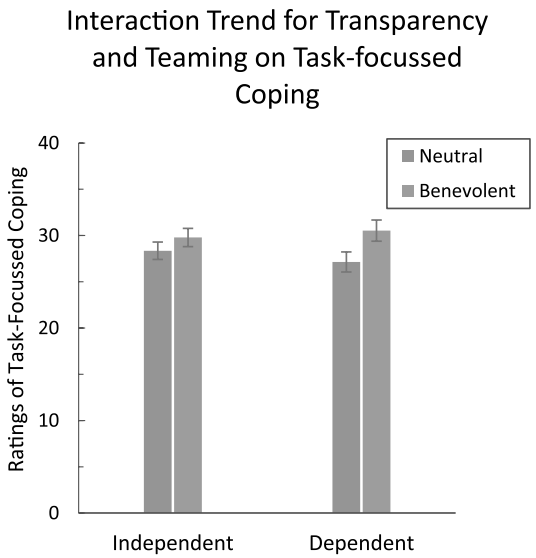


Figure 6. Trend for interaction between teaming and transparency on reported task-focused coping ($p = .056$).
Note. Standard error bars shown.

neutral condition ($M = 40.94$). The Transparency × Scale interaction was nonsignificant, implying a uniform effect across the different ratings.

Trust Measures

To test H3 for perceived trustworthiness, a 2 (Team Type: Independent/Dependent) × 2 (Transparency: Neutral/Benevolent) × 3 (Scale: Ability/Integrity/Benevolence) mixed-model ANOVA was run on the adapted TTP measure. No significant main effects or interactions were found.

To test H2 for trust, a 2 (Team Type: Independent/Dependent) × 2 (Transparency: Neutral/Benevolent) × 3 (Scale: Reliability/Technical Confidence/Understandability) mixed-model ANOVA was run on the HCT scores. The HCT showed a significant main effect for team type, $F(1, 37) = 4.12$, $p = .05$, $\eta_p^2 = .10$, where individuals dependent on the teammate ($M = 3.591$) reported higher trust than individuals performing independently of the Viper 2 ($M = 3.208$). In addition, there was a main effect for scale, $F(1.63, 63.45) = 7.15$, $p = .01$, $\eta_p^2 = .16$. Paired t tests were Bonferroni-corrected (three comparisons; critical $p = .017$) to reveal that Perceived Understandability ($M = 3.55$) was significantly greater than Perceived Reliability ($M = 3.26$), but no different from Perceived Technical Competence ($M = 3.39$).

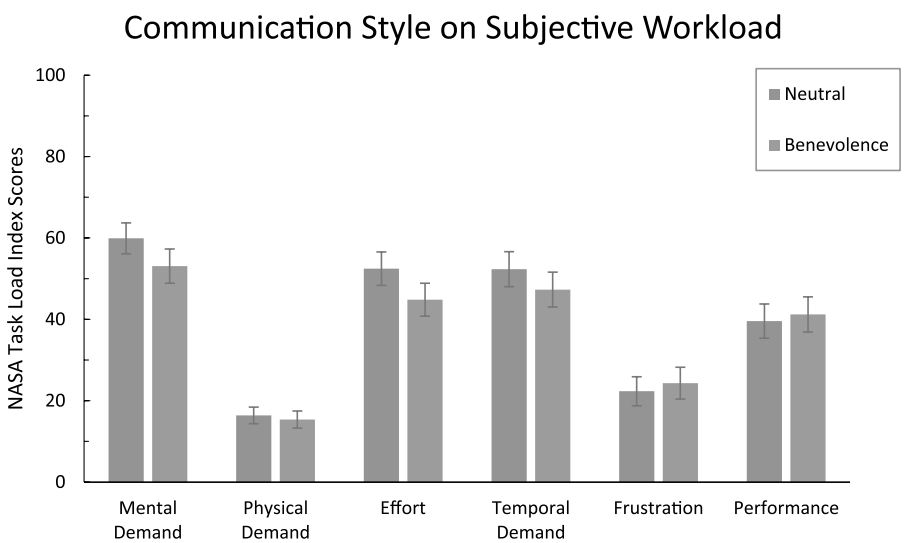


Figure 7. Interaction trend between transparency and scale on TLX scores ($p = .07$).
Note. Main effects were significant for communication style ($p < .05$) and TLX scores ($p < .001$). Standard error bars shown. Performance is scored out of 100, where 100 = higher performance. TLX = task load index.

Technology Acceptance of Machines

To test H3, a 2 (Team Type: Independent/Dependent) × 2 (Transparency: Neutral/Benevolent)) × 5 (Scale: Competence/Integrity/Usefulness/Trust/Intention to Adopt) mixed-model ANOVA was performed. Analyses revealed a significant main effect of main effect for Scale, $F(2.64, 92.33) = 7.04$, $p < .001$, $\eta_p^2 = .26$. Paired t tests (three comparisons; critical $p = .017$) were Bonferroni-corrected, revealing that the measure for Trust was significantly lower ($M = 4.47$) than Perceived Usefulness ($M = 4.75$) and Integrity ($M = 4.82$). Intention to Adopt ($M = 5.09$) was significantly higher than Perceived Usefulness ($M = 4.75$) and Integrity ($M = 4.82$). There was also a significant main effect for Team Type, $F(1, 35) = 4.59$, $p < .05$, $\eta_p^2 = .12$. Overall acceptance of machines was rated higher in the dependent condition ($M = 5.12$) than in the independent condition ($M = 4.32$).

Decision Confidence

To test H4, a 2 (Team Type: Independent/Dependent) × 2 (Transparency: Neutral/Benevolent) × 2 (Scale: Confidence in Self/Confidence in Robotic Partner Decision) mixed ANOVA was run on the average of confidence ratings measured after every trial (12 per condition). Confidence scales revealed a significant main effect, $F(1, 40) = 9.92$, $p < .01$. Confidence in self was higher ($M = 77.35$) than in partner ($M = 70.90$).

DISCUSSION

The current study examined the joint impact of transparency and partnership type on various team-related, trust and state factors in a simulated military flight mission. The overarching aim was to explore the potential benefits for HMT of benevolent social communication in a military-level operation. In this context, the machine's communication of support was intended as a form of transparency of intent that provides the human with insight into the machine's goals. Expressions of benevolence may facilitate positive states and building shared mental models of team coordination between teammates, especially when their partnership requires human dependence on the machine. Therefore, investigating the benefit of communicating intent to support was believed to provide further design considerations for HMT in military settings.

Findings confirm that design of robot-to-human communications is important for optimizing operator perceptions of teaming and workload. Feelings of teaming and trust were increased by transparency and teaming context, respectively. The findings support the notion that individuals are inclined to team with an AT (Nass et al., 1996) in a military task. In addition, the study provides support for using communication as a method of transparency into both functional and teaming intent of the system. Most importantly, the use of transparency, as it is operationalized in this study, reduces subjective workload experienced by the human partner.

There were effects on outcomes for both the transparency and team type manipulations, but hypotheses were only partially supported. Table 1 summarizes where support was found. It suggests that impacts of benevolence may be more nuanced than anticipated, and there was a dissociation between outcomes sensitive to transparency and to team dependence. Benevolence was found to enhance certain aspects of team perceptions, workload, and trust as expected, but other outcome measures were unaffected. The effects of benevolence seem rather selective. From a methodological standpoint, this result points to the need to use measures of appropriate sensitivity in HMT research. From a theoretical perspective, it indicates the need for a more fine-grained theory of how benevolence influences teaming. In the remainder of this discussion, we identify the measures most sensitive to the manipulation, examining each hypothesis in turn. We also indicate study limitations.

Aspects of Teaming Sensitive to Benevolence

The hypothesized effect of transparency on perception of teaming (H1) was partially supported. Benevolent communication increased ratings of team collaboration, averaged across teaming type. There was no effect on team cohesion, trust, or efficacy, possibly reflecting the nature of the task. The task was simple and

TABLE 1: Summary of Support for Hypotheses

Hypothesis	Outcome: Core Construct(s)	Outcome: Measure(s)	Level of Support
H1	Teaming perceptions	Collective efficacy, team collaboration, team trust, group cohesion	Partial: supported for team collaboration
H2	Workload and stress	Workload, stress state, coping	Partial: supported for workload
H3	Trust in AT	Trustworthiness and trust, technology acceptance, HCT	Partial: supported for HCT trust
H4	Confidence	Confidence in decision and partner	Unsupported

Note. AT = autonomous teammate; HCT = human–computer trust.

feedback was not given after each trial; therefore, perceptions of how well the team was working together could not shift. In a future study where multiple entities require coordination between the AT and human, success can be monitored and perceived team efficacy will be based on that success. In addition, interpersonal trust in team members may not be appropriate for the current task (Naquin & Paulson, 2003) as the AT did not attempt to "get the upper hand" or "exploit the situation." The absence of an effect on this trust scale points to the importance of using appropriate measures of trust for the context and in future research.

H2 was tested to determine any beneficial effects of transparency, when teammates were dependent on one another, on subjective stress and workload. Surprisingly, there was no effect of transparency or team type on any of the stress factors, although a tendency toward increased task-focused coping just missed statistical significance. Of great interest to the HMT community was the reduction of workload found in the Benevolent condition. Findings from studies by Selkowitz et al. (2015) and Mercado et al. (2016) revealed that added transparency information did not mitigate workload. Transparency in previous work was provided within a SA model (see Barnes et al., 2017, for review) where visual displays were used to provide transparency into the system's current state (Level 1), environmental constraints and logic (Level 2), and finally its projected state (Level 3). It is possible that the difference in modality for transparency drove the reduction in reported workload. The use of audio messages may be naturally less taxing than additional display information, as the primary task was visual in nature.

The effect of transparency on trust (H3) yielded mixed results: teaming type affected trust assessed with the HCT (Madsen & Gregor, 2000), but not measures of trustworthiness or technology acceptance. Trust, averaged across technical confidence, understandability, and reliability, was greater when participants were dependent on the AT to assist with their task of jamming with the SAM. Trust when the participant and AT were performing separate tasks was lower, perhaps because there is no need to trust the AT; it performed a surveillance task which had no way of affecting the participant's task and so there was no vulnerability. The main effect of scale in the HCT measure indicated that perceived understandability and technical competence were stronger drivers of trust than was reliability in this task. This may be driven by the fact that the AT's waypoint errors harmed perceptions of reliability overall but that understanding its purpose was straightforward and its ability to perform its main function did not waver.

Technology Acceptance (Ghazizadeh et al., 2012) was also analyzed as part of H2, and the main effect for scale revealed that intention to adopt was the strongest factor in technology acceptance. This scale's items indicated a willingness to use the system. High ratings for Integrity and Perceived Usefulness support findings in the HCT measure; however, mean trust from the TAM scale was surprisingly low. These TAM

items may have requested ratings for an interpersonal trust that either did not exist in this task or did not fit the participant's thoughts of an AT (i.e., my partner puts my interests first). Trust of this kind may be important in a more complex teaming task where the human partner must carry out multiple tasks, requiring consideration, understanding, and backup from the AT.

Transparency was expected to affect confidence in partner (H4) but it had no effect. Generally, individuals reported more confidence in their decision, whether to continue or abort, than confidence in the AT. This finding is interesting when taken together with the HCT findings; although individuals are not necessarily confident in the AT, they are willing to trust it. Typically, high self-confidence is associated with low trust (see Hoff & Bashir, 2015) and so the current finding indicates that there may be additional elements for confidence in the AT that were not captured in the single confidence question. Perhaps the question should have been separated into confidence in flight ability versus confident in radar capability as these were the only aspects of the AT to make judgments on. Future research should break down confidence into more fine-grained constructs based on the complexity of the task.

Limitations

Any allusion to the benefit of benevolence seems to imply a reduction of uncertainty, which in turn may mitigate stress and workload. However, the role of uncertainty was not measured directly. Similarly, it may have been beneficial to assess SA (Endsley, 2015), that is, the operator's awareness of current state and future state. SA maps onto some models of trust (Barnes et al., 2017) and its assessment would more definitively relate the positive effects of sharing intent to support to the reduction of uncertainty of future outcomes. However, the simplicity of the current task would not have benefited from SA assessments as the only SA needed is to know whether the system is staying on path. Future studies may be able to look at the impact of transparency of intent in a more dynamic task environment where SA can be assessed.

In addition, the communication of errors may have had a negative effect on trust, as measured by the team trust items. These items point to interpersonal trust and feelings that the system was unreliable and untruthful. Perhaps the information was useful for team processing but did hurt trust. Future work may remedy this by communicating Level 2 information (Barnes et al., 2017) and could still yield a benefit to subjective workload but also an increase in trust.

Results may also be specific to the specific teaming context and population examined. The mental models for HMT the human brings to the mission may vary with contextual factors (Kiesler & Goetz, 2002). For example, in the present Air Force context, participants may have been disposed to trust the AT because it is unlikely that the Air Force would deploy an ineffective or unreliable system. The community sample of nonpilots may also have a distinctive set of beliefs about automated systems; results may have been different in samples of pilots, students, or information technology experts. Individual differences in aspects of mental models such as perfectionistic expectations (Merritt, Unnerstall, Lee, & Huber, 2015) may also moderate the impact of manipulations of robot communications. Larger sample sizes would also be desirable in future research to determine whether trends observed for subjective state and coping research would be significant with greater statistical power.

Generalization of findings to real-world military contexts is limited by the simplicity of interaction with the AT as well as the use of a naïve population. The current study was aimed to carefully control for the complications of a real-world scenario to observe the subjective responses of normal adults to an AT in a military environment. Other studies in HMT observe responses in safer, less-threatening environments such as financial decision-making (de Visser et al., 2016). Although further study is necessary to investigate the impact of workload on individuals with flight experience, the study is the first of its kind to determine whether social dynamics is useful in such a context.

CONCLUSION

As research and development of robotic partners advances, consideration of the social aspects relevant to teaming may be useful in

understanding team effectiveness across multiple domains and designing beneficial transparency. The amount of social cueing may differ across applications, but communication of intent to support may still be beneficial in teams performing in high stress environments such as military applications. However, both benevolence and team interdependence had somewhat subtle effects and further work is needed to develop underlying theory and system design to optimize communication. What is also still needed is consideration of baseline expectations for socialization; for example, can we train teaming schemas to remove negative biases toward robotic partners? In addition, there remains a question of how high tempo tasks interact with transparency of teaming intentions. Although the current study was simple in its task requirements, it did reveal that monitoring the actions of a teammate is taxing and the burden is reduced by transparency of support intentions.

The study provides the practitioner further evidence for the importance of design for machine-to-human communication, extending existing work on situational awareness and workload (Chen & Barnes, 2014) and the impact of scheduling of messages (Dorneich et al., 2012; Gombolay, Blair, Huang, & Shah, 2017). The current study demonstrates that communication can be designed to convey a social message that regulates teaming processes via affect management. As technology enhances machine capabilities for comprehension and communication, more consideration of machine messaging and its intent will be necessary for HMT. Designers also need to consider how to support human-to-machine communication (Mavridis, 2015) that regulates machine messaging, so that, for example, benevolent messages can be delivered when they will be perceived as supportive and task-relevant, rather than distracting. Effective two-way communication is the key to seamless human–machine interaction akin to cohesive human–human teaming.

ACKNOWLEDGMENT

This project was supported by funds provided by the Air Force Office of Scientific Research, Grant Award No. 16RHCOR367.

ORCID iD

April Rose Panganiban https://orcid.org/0000-0001-6922-4086

REFERENCES

Barnes, M. J., Chen, J. Y., & Hill, S. (2017). *Humans and autonomy: Implications of shared decision making for military operations* (No. ARL-TR-7919). Aberdeen Proving Ground, MD: U.S. Army Research Laboratory.

Chen, J. Y., & Barnes, M. J. (2014). Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems*, *44*(1), 13–29.

Cheung, C. M., & Lee, M. K. (2001). Trust in internet shopping: Instrument development and validation through classical and modern approaches. *Journal of Global Information Management*, *9*(3), 23–35.

De Jong, B. A., Dirks, K. T., & Gillespie, N. (2016). Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *Journal of Applied Psychology*, *101*, 1134–1150.

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*, 331–349.

Dorneich, M. C., Ververs, P. M., Mathan, S., Whitlow, S., & Hayes, C. C. (2012). Considering etiquette in the design of an adaptive system. *Journal of Cognitive Engineering and Decision Making*, *6*, 243–265.

Driskell, J. E., Salas, E., & Driskell, T. (2018). Foundations of teamwork and collaboration. *American Psychologist*, *73*, 334–348.

Endsley, M. R. (2015). Situation awareness misconceptions and misunderstandings. *Journal of Cognitive Engineering and Decision Making*, *9*(1), 4–32.

Ghazizadeh, M., Peng, Y., Lee, J. D., & Boyle, L. N. (2012, September). Augmenting the technology acceptance model with trust: Commercial drivers' attitudes towards monitoring and feedback. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*, 2286–2290.

Gombolay, M., Blair, A., Huang, C., & Shah, J. (2017). Computational design of mixed-initiative human–robot teaming that considers human factors: Situational awareness, workload, and workflow preferences. *The International Journal of Robotics Research*, *36*, 597–617.

Groom, V., & Nass, C. (2007). Can robots be teammates? Benchmarks in human–robot teams. *Interaction Studies*, *8*, 483–500.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology, Vol. 52. Human mental workload* (pp. 139–183). Oxford, UK: North-Holland.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*, 407–434.

Ilies, R., Johnson, M. D., Judge, T. A., & Keeney, J. (2011). A within-individual study of interpersonal conflict as a work stressor: Dispositional and situational moderators. *Journal of Organizational Behavior*, *32*, 44–64.

Kiesler, S., & Goetz, J. (2002, April). Mental models of robotic assistants. In *CHI'02 extended abstracts on human factors in computing systems* (pp. 576–577). New York, NY: ACM.

Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a "team player" in joint human-agent activity. *IEEE Intelligent Systems*, *19*(6), 91–95.

Komiak, S. Y., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly*, *30*, 941–960.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*, 50–80.

Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., & Paiva, A. (2013). The influence of empathy in human–robot relations. *International Journal of Human-Computer Studies*, *71*, 250–260.

LePine, J. A., Piccolo, R. F., Jackson, C. L., Mathieu, J. E., & Saul, J. R. (2008). A meta-analysis of teamwork processes: Tests of a multidimensional model and relationships with team effectiveness criteria. *Personnel Psychology*, *61*, 273–307.

Li, X., Hess, T. J., & Valacich, J. S. (2008). Why do we trust new technology? A study of initial trust formation with organizational information systems. *The Journal of Strategic Information Systems*, *17*, 39–71.

Lyons, J. B. (2013, March). *Being transparent about transparency: A model for human-robot interaction*. In D. Sofge, G. J. Kruijff, & W. F. Lawless (Eds.), *AAAI spring symposium series* (pp. 48–53). Menlo Park, CA: AAAI Press.

Lyons, J. B., & Guznov, S. Y. (2018). Individual differences in human-machine trust: A multi-study look at the perfect automation schema. *Theoretical Issues in Ergonomics Science*, *20*, 440–458.

Lyons, J. B., Ho, N. T., Fergueson, W. E., Sadler, G. G., Cals, S. D., Richardson, C. E., & Wilkins, M. A. (2016). Trust of an automatic ground collision avoidance technology: A fighter pilot perspective. *Military Psychology*, *28*, 271–277.

Lyons, J. B., Sadler, G. G., Koltai, K., Battiste, H., Ho, N. T., Hoffmann, L. C., . . . Shively, R. (2017). Shaping trust through transparent design: Theoretical and experimental guidelines. In P. Savage-Knepshield & J. Chen (Eds.), *Advances in human factors in robots and unmanned systems: Proceedings of the AHFE 2016 international conference on human factors in robots and unmanned systems* (pp. 127–136). Cham, Switzerland: Springer.

Madsen, M., & Gregor, S. (2000, December). Measuring human-computer trust. In *11th Australian conference on information systems* (Vol. 53, pp. 6–8).

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.

Marlow, S. L., Lacerenza, C. N., Paoletti, J., Burke, C. S., & Salas, E. (2018). Does team communication represent a one-size-fits-all approach? A meta-analysis of team communication and performance. *Organizational Behavior and Human Decision Processes*, *144*, 145–170.

Matthews, G., & Campbell, S. E. (1998, October). Task-induced stress and individual differences in coping. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *42*, 821–825.

Matthews, G., Campbell, S. E., Falconer, S., Joyner, L. A., Huggins, J., Gilliland, K., . . . Warm, J. S. (2002). Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. *Emotion*, *2*, 315–340.

Matthews, G., Wohleber, R. W., & Lin, J. (2019). Stress, skilled performance, and expertise: Overload and beyond. In P. Ward, J. Maarten Schraagen, J. Gore, & E. M. Roth (Eds.), *The Oxford handbook of expertise* (pp. 1–39). New York, NY: Oxford University Press.

Mavridis, N. (2015). A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, *63*, 22–35.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integration model of organizational trust. *Academy of Management Review*, *20*, 709–734.

Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human–agent teaming for multi-UxV management. *Human Factors*, *58*, 401–415.

Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring individual differences in the perfect automation schema. *Human Factors*, *57*, 740–753.

Naquin, C. E., & Paulson, G. D. (2003). Online bargaining and interpersonal trust. *Journal of Applied Psychology*, *88*(1), 113–120.

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, *45*, 669–678.

Pfaff, M. S. (2012). Negative affect reduces team awareness: The effects of mood and stress on computer-mediated team communication. *Human Factors*, *54*, 560–571.

Riggs, M. L., & Knight, P. A. (1994). The impact of perceived group success-failure on motivational beliefs and attitudes: A causal model. *Journal of Applied Psychology*, *79*, 755–766.

Rozell, E. J., & Gundersen, D. E. (2003). The effects of leader impression management on group perceptions of cohesion, consensus, and communication. *Small Group Research*, *34*, 197–222.

Schaefer, K. E., & Scribner, D. R. (2015, September). Individual differences, trust, and vehicle autonomy: A pilot study. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*, 786–790.

Selkowitz, A., Lakhmani, S., Chen, J. Y., & Boyce, M. (2015, September). The effects of agent transparency on human interaction with an autonomous robotic agent. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*, 806–810.

Smith-Jentsch, K. A., Kraiger, K., Cannon-Bowers, J. A., & Salas, E. (2009). Do familiar teammates request and accept more backup? Transactive memory in air traffic control. *Human Factors*, *51*, 181–192.

Wang, W., & Benbasat, I. (2005). Trust in and adoption of online recommendation agents. *Journal of the Association for Information Systems*, *6*(3), 72–101.

Wynne, K. T., & Lyons, J. B. (2018). An integrative model of autonomous agent teammate-likeness. *Theoretical Issues in Ergonomics Science*, *19*, 353–374.

April Rose Panganiban received her Ph.D. (2013) in Human Factors Psychology from the University of Cincinnati, Cincinnati, OH. She has worked as a contractor and now, civilian Research Psychologist for the Air Force Research Laboratory since 2007. Her research interests include human-machine teaming and the contribution of human state changes (e.g., stress, affect and workload), individual differences and personality to this teaming dynamic.

Gerald Matthews received B.A. (1980) and Ph.D. (1984) degrees in experimental psychology from the University of Cambridge, Cambridge, U.K. He has

held faculty positions at Aston University, the University of Dundee, and the University of Cincinnati. In 2013 he took up his current position as Research Professor at the Institute for Simulation and Training, University of Central Florida, Orlando. His research interests include human-robot interaction, human factors in transportation and unmanned systems, and the impacts of stress, workload and fatigue on human performance.

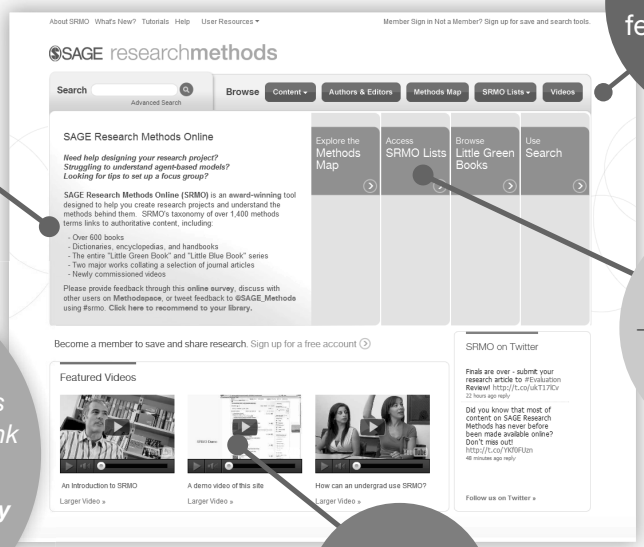
Michael D. Long received a B.S. (2016) in psychology from the University of Cincinnati, Cincinnati, OH. He is currently enrolled at the University of Cincinnati as a student for a B.S. in neuropsychology, and he is expecting to graduate in spring 2020. He currently works as a contractor for the Air Force Research Laboratory through Oak Ridge Institute for Science and Education (ORISE) since 2016.

SAGE researchmethods

The essential online tool for researchers from the world's leading methods publisher

From basic explanations to advanced discussion, **SAGE Research Methods** will lead you to the content you need

More content and new features added this year!



Discover **Methods Lists** – methods readings suggested by other users

"I have never really seen anything like this product before, and I think it is really valuable."

John Creswell, University of Nebraska–Lincoln

Watch video interviews with leading methodologists

Explore the **Methods Map** to discover links between methods

Based on a custom-designed taxonomy with over 1,400 qualitative, quantitative, and mixed methods terms

More than 120,000 pages of book, journal, and reference content to support your learning

find out more at **www.sageresearchmethods.com**

JUST PUBLISHED!

Usability Assessment: How to Measure the Usability of Products, Services, and Systems

Volume 1, Users' Guides to Human Factors and Ergonomics Methods

By Philip Kortum

Usability Assessment is a concise volume for anyone requiring knowledge and practice in assessing the usability of any type of product, tool, or system *before* it is launched. It provides a brief history and rationale for conducting usability assessments and examples of how usability assessment methods have been applied, takes readers step by step through the process, highlights challenges and special cases, and offers real-life examples. By the end of the book, readers will have the knowledge and skills they need to conduct their own usability assessments without requiring that they read textbooks or attend workshops.

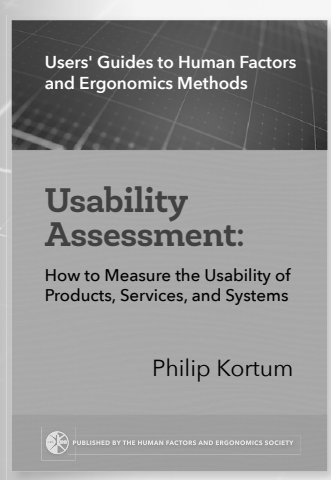


Table of Contents

1. What Is Usability Assessment?
2. Why Assess Usability?
3. Prepare to Perform the Usability Evaluation
4. Create Your Test Plan
5. Perform the Usability Test
6. Special Cases of Usability Assessment
7. Real-Life Example 1: Corporate Web Portal
8. Real-Life Example 2: High-Security Voting System
9. Some Parting Advice

This book will be valuable for undergraduate and graduate students; practitioners; usability professionals; human-computer interaction professionals; researchers in fields such as industrial design, industrial/organizational psychology, and computer science; and those working in a wide range of content domains, such as health care, transportation, product design, aerospace, and manufacturing.

ISBN 978-0-945289-49-4
120 pp., 7" x 10" paperback and e-book
http://www.hfes.org/publications/

PUBLISHED BY THE HUMAN FACTORS AND ERGONOMICS SOCIETY